Experiments With Assessment of Creative Systems: An Application of Ritchie's Criteria

Francisco C. Pereira[1], **Mateus Mendes**[2], **Pablo Gervás**[3], **and Amílcar Cardoso**[1]
[1] Dep. de Eng. Informática da FCTUC, Coimbra, Portugal
[2] ESTGOH, Oliveira do Hospital, Portugal
[3] Universidad Complutense de Madrid, Spain
{camara,amilcar}@dei.uc.pt, mmendes@estgoh.ipc.pt, pgervas@sip.ucm.es

Abstract

Creativity assessment is an area in demand of research, for creativity is a recent field of study and its principles are not well understood. Ritchie's Criteria are a recent proposal, consisting of fourteen principles to assess the creativity of computer programs. We have applied those criteria to three different systems: Wasp, a poem generator, Divago, a conceptual blender, and Dupond, a sentence paraphraser. The results are hereby discussed and compared, and the main difficulties of applying this methodology are pointed out.
Keywords: Ritchie's Criteria, Assessing Creativity, Wasp, Divago, Dupond.

1 Introduction

The assessment of the creativity of a computational system (or any other) is understandably one of the most important challenges in the area of Computational Creativity (CC). The problem comes from the definition of *creativity* itself, which is even more controversial than that of *intelligence* within AI. While creativity is often a mystified and passionate subject, it is a task of CC to understand and formalize it at the achievable limits. So far, two fundamental aspects are almost consensual: something to be called creative needs to escape the trail of predictability (an aspect often connected to *novelty*, *non-typicality*, *surprise*); and a creative product must be of value (an aspect often connected to *usefulness*, *purposefulness*). Departing from such aspects and understanding the need of an emerging field, Graeme Ritchie proposed a set of fourteen criteria to assess the creativity of a system.

In this paper, we bring together three independent creativity analyses made for three different systems. More than comparing the systems or claim for their creativity, the goals are to present the processes followed for the assessment, the problems found, and the ideas that may have emerged. Thus, we discuss the strengths and weaknesses found in Ritchie's criteria, contributing to those that intend to apply these criteria or even propose improvements to the framework.

The three systems involved are: Wasp (a poetry generation system); Divago (a concept invention system) and Dupond (a paraphraser system). In some way, all three systems deal with language processing. Both Wasp and Dupond produce natural language outputs, while Divago generates semantic networks. The latter has an internal validation mechanism which enables it to estimate the *novelty* and *usefulness* of the outputs with respect to what is known in the Knowledge Base. The former two were subject of enquiries given to people to classify their outputs.

The next section of this paper will give the reader a short overview of Ritchie's criteria. We strongly advise the reader unaware of this work to read it, since space restrictions will not allows us to get into great detail. Section 3 will describe the systems and the experiments made. The discussion is taken up in section 4 and the paper ends with some conclusions (section 5).

2 Ritchie's Criteria

Ritchie proposes a set of criteria for assessing creativity on the basis of the results of the system (i.e. the product), its initial data and the items that gave rise to its construction (the *inspiring set*) [Ritchie, 2001]. Prior to describing the criteria, we have to give a set of definitions: **B** - Basic item - an entity that a program produces. "This is *not* a definition of what would count as successful or valid output for the program, merely a statement of the data type it produces"; **I** - the inspiring set - the set of basic items that implicitly or explicitly drove the development of the program; **R** - the set of results produced by the system; **typ** - typicality of the items; **val** - value of the items.

Ritchie proposes fourteen criteria to assess the creativity of a system's output, R. Although it is assumed that R corresponds to the result(s) of a single run, the generalization of these criteria to a set of runs is also suggested, in order to cover the general behaviour of the system. The criteria intend to measure the behaviour of the system in terms of average (AV) quality of results, their typicality and of their ratios ($ratio$), with regard to R and to the set of typical and valued items.

The criteria are summarised in table 1.

2.1 Followups

Measuring novelty and quality

Pease, Winterstein and Colton [Pease *et al.*, 2001] suggest methods that take into account issues like complexity of the search space, difference to archetypes and inspiring sets,

Crit.	Formalization	Informal meaning
1	$AV(typ, R) > \theta_1$	Average typicality
2	$ratio(T_{\alpha,1}(R), R) > \theta_2$	Ratio typical results / all results
3	$AV(val, R) > \theta_3$	Average quality
4	$ratio(V_{\gamma,1}(R), R) > \theta_4$	Ratio good results / all results
5	$ratio(V_{\gamma,1}(R) \cap T_{\alpha,1}(R), T_{\alpha,1}(R)) > \theta_5$	Ratio good typical results / all results
6	$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), R) > \theta_6$	Ratio good atypical results / all results
7	$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), T_{0,\beta}(R)) > \theta_7$	Ratio good atypical results / atypical results
8	$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), V_{\gamma,1}(R) \cap T_{\alpha,1}(R)) > \theta_8$	Ratio good atypical results / good typical results
9	$ratio(S_{\mathcal{B}}(typ, val) \cap R, S_{\mathcal{B}}(typ, val)) > \theta_9$	Ratio results in the inspiring set / inspiring set
10	$ratio(R, S_{\mathcal{B}}(typ, val) \cap R) > \theta_{10}$	Ratio all results / results in the inspiring set
11	$AV(typ, (R - S_{\mathcal{B}}(typ, val))) > \theta_{11}$	Average typicality of new results
12	$AV(val, (R - S_{\mathcal{B}}(typ, val))) > \theta_{12}$	Average quality of new results
13	$ratio(T_{\alpha,1}(R - S_{\mathcal{B}}(typ, val)), R) > \theta_{13}$	Typical new results / new results
14	$ratio(V_{\gamma,1}(R - S_{\mathcal{B}}(typ, val)), R) > \theta_{14}$	Good new results / result

Table 1: Summary of the fourteen criteria (for suitable θ_i, α, β and γ).

among others. While some measures can be directly applied (e.g. complexity of the space, novelty), others are extremely difficult, especially the ones dependent on the notion of *bag*. According to the authors, a *bag* contains all pieces of knowledge in the program P, which contribute to the generation and evaluation of x: $H_x = [k \in P : k$ is used to generate $x \in R]$. In systems where the generation is based on stochastic procedures (e.g. Divago's GA), it becomes hard to determine exactly what contributed to the generation of an outcome. Perhaps this kind of analysis must be considered in the design of the system (e.g. to include a *bag* tracking mechanism) rather than post-hoc (as happened with Divago).

The effect of input knowledge

One of the main problems in evaluating computational creativity (and of AI systems in general) relates to the extent to which the system's knowledge is *fine-tuned*, i.e. the system essentially replicates known items to a greater extent than it causes the generation of novel high-valued items. Colton, Pease and Ritchie [Colton *et al.*, 2001] propose a set of criteria for evaluating the effect of input knowledge. They first compare the behaviour of a system with and without an input knowledge K in terms of the quality of the output (highly valued, reinventions and others). From the sets obtained, they estimate the fine-tuning of a program P using input knowledge K, depending on the set of non-replicated high valued items created using K (in the limit, if K only leads to replications, then P using K is completely fine-tuned).

Considering the systems described in this paper, only Divago has been subject to such analysis. The author tested with a set of *frames* and configurations, reaching the conclusion that while some frames (i.e. some input knowledge K) lead to more replications, one cannot talk about fine-tuning (we will see below that Divago has few reinventions) in Divago. Conversely, from the same analysis, Divago has demonstrated to be highly unpredictable (which in some situations implies a somewhat undesirable lack of control).

3 Experiments

3.1 Wasp

The System

The WASP system [Gervás, 2000] draws on prior poems and a selection of vocabulary provided by the user to generate a metrically driven re-combination of the given vocabulary according to the line patterns extracted from prior poems.

The *inspiring set* is taken to be a specific 16th century Spanish classical sonnet. This establishes a number of restrictions on the poetry that is to be composed. Lines should have 11 syllables, according to very strict stress patterns.

To simplify matters, the artifacts that the program will aim for will be samples in isolation of one of the simpler stanzas that make up a sonnet: a *cuarteto*, a stanza of four lines rhyming ABBA.

The construction process that is employed is designed to ensure that all resulting items have the correct syllable count and a valid pattern of stressed syllables for each line. Given a specific stanza to aim for, the system attempts to build an instance of this stanza based on the set of line patterns it receives and the available vocabulary. Wherever several possible choices of words match the metric constraints, the program makes a random choice. This provides the non-determinism required to obtain multiple results on different runs.

Results

Each run of the program with such an initialisation produces either a complete stanza of the desired form or as many lines as can be produced while meeting the metric criteria. In order to obtain results that can be analysed according to Ritchie's framework, each set of 12 runs with the same initialisation is studied as a single set of results. Fourteen different initialisations are considered. This gives a total of 168 resulting poems. Each poem is evaluated by a team of volunteers, who are asked to provide two numerical values: one measuring the syntactic correctness of the poem (on a scale from 0 to 5) and one measuring the aesthetic qualities of the poem (on a similar scale). These values are combined with the number

Crit.	Meaning	Experiments				
		$\alpha = 0.7$ $\gamma = 0.7$	$\alpha = 0.5$ $\gamma = 0.5$	$\alpha = 0.3$ $\gamma = 0.3$	$\alpha = 0.7$ $\gamma = 0.3$	$\alpha = 0.3$ $\gamma = 0.7$
1	Average typicality	0.71	0.71	0.71	0.71	0.71
2	Typical results / results	0.54	0.88	0.89	0.54	0.89
3	Average quality	0.47	0.47	0.47	0.47	0.47
4	Good results / results	0.24	0.50	0.68	0.68	0.24
5	Good typical results / results	0.36	0.57	0.77	0.89	0.28
6	Good atypical results / results	0.05	0.00	0.00	0.21	0.00
7	Good atypical results / atypical results	0.12	0.00	0.00	0.45	0.00
8	Good atypical results / good typical results	0.28	0.00	0.00	0.44	0.00
9	Results in the inspiring set / inspiring set	0.00	0.00	0.00	0.00	0.00
10	Results / results in the inspiring set	N/A	N/A	N/A	N/A	N/A
11	Average tipicallity of new results	0.71	0.71	0.71	0.71	0.71
12	Average quality of new results	0.47	0.47	0.47	0.47	0.47
13	Typical new results / new results	0.54	0.54	0.54	0.54	0.54
14	Good new results / results	0.24	0.24	0.24	0.24	0.24

Table 2: Results of Ritchie's Criteria for WASP's experiments, considering $\alpha = \beta$, val assigned by volunteer evaluators and typ worked out from number of lines and syntactic coherence.

of lines of each poem to provide an approximation to the two evaluation functions required.

The application of Ritchie's criteria to the results of this generation process are presented in detail in [Gervás, 2002]. A summary of the values resulting from this application is presented in table 2.

A poem is considered typical if it has the required number of lines and it has a syntactically correct reading. The value obtained for *typ* is actually the result of combining mathematically the values assigned for syntactic correctness and the number of lines obtained for each attempted instance of the stanza. The actual formula applied to obtain the final value corresponds to what Ritchie calls a *weighted property rating scheme*, as used for evaluating typicality. The role of the weights employed in the actual combination is discussed in detail [Gervás, 2002].

A poem is considered good depending on the value assigned to its aesthetic qualities by the evaluators.

What the Criteria Say about WASP

Only the first eight criteria are relevant, because none of the inspiring set reappears in the result. This is apparent in the fact that criterion 10 tends to infinity as the number of results already present in the inspiring set tends to 0. This is due to the fact that the construction process actually first factorises and then recombines elements of the inspiring set, adding additional words from the vocabulary. This reduces greatly the probability that an element in the inspiring set be generated anew by the system. An immediate consequence is that criterion 9 drops to zero and criterion 10 runs up to infinity. Additionally, those criteria designed to capture specific differences between items that are new and items already in the inspiring set produce the same score as the original criteria they are evolved from (the same values result for criteria 11 and 1, 12 and 3, 13 and 2, 14 and 4).

A question that may need detailed discussion is how one identifies whether an element in the inspiring set is reappearing in the results. For this version of the system, none of the *cuartetos* in the inspiring set appears as such among the results, but some of the lines of the poems in the inspiring set may reappear, and—given the construction procedure employed—all of the lines in the results will have a syntactic structure that is borrowed from the lines in the inspiring set.

The system is better at producing typical items than at producing good items (score higher for criterion 1 than for criterion 3) and higher for criterion 2 than for criterion 4. This makes sense, since all system decisions (algorithms applied and constraints imposed) during the construction process are concerned with ensuring the production of typical items, rather than good ones. In fact, the system has no means for identifying good items, and therefore cannot be expected to aim towards them during construction.

Atypical results score badly in terms of quality. This may be due to evaluators not having a clear idea of whether their judgement on the quality should take into account how typical the item is. Evaluators may be awarding good scores on quality to items that are typical. This would imply that their own reaction is to apply criterion 5 rather than criterion 4. The fact that the system performs better under criterion 5 than criterion 4 with these evaluators may be taken as evidence in favour of this interpretation. Criterion 8 provides an indication of this relation (low presence of atypical results among the good results).

The Choice of Threshold Parameters

The three threshold values introduced by Ritchie (α, β and γ) are applied to distinguish highly rated items whether on typicality or quality. Of these, α and β were kept equal to one another throughout the set of experiments, indicating that items that did not rate highly on typicality have been considered atypical. A finer grained approach would establish a low threshold below which items would be considered as atypical. This might establish a high threshold value to determine when an item is typical and a low threshold value to deter-

mine when an item is atypical.

Experiments were carried out varying the thresholds that are used to distinguish highly rated items in each class (typical or good). These correspond to the different columns in table 2. Criteria 1 and 3 are not affected by this change, since they do not refer to the threshold value. Therefore they are omitted from the following discussion.

The threshold value on quality determines how many items are considered good, and therefore affects criteria 4 through to 8. The threshold value on typicality affects criterion 2 and criteria 5 through to 8.

It can be seen from the results that lowering the typicality threshold results in a zero score for criteria 6 to 8. This is because they involve good atypical results. By lowering the typicality threshold the number of atypical items is reduced, and any reduction brings down the number of good items to be found among them. Criteria 2 (regarding typicality) and 4 (concerned with quality) are inversely proportional to the threshold applied in each case—the value for the corresponding criteria falls when the threshold rises and falls when it rises. Criteria 5 is different in every case because it involves both thresholds.

There is a great variation between the values obtained for each of these criteria when the thresholds are moved. This implies that the assignment of specific values for these thresholds should be established beforehand based on domain specific criteria, or oriented towards the specific aims that have been established for the system.

3.2 Divago

The System

Divago [Pereira, 2005] is a system for Concept Invention that aims to generate concepts via a mechanism of *Conceptual Blending* [Fauconnier and Turner, 1998]. In Divago, a *concept* is defined by a micro-theory which comprises two levels of knowledge: the concept map—a semantic network which relates the concept to other concepts; the frames, rules and integrity constraints, which describe patterns associated to the concept (a *bus* fits the pattern of *transport means*) as well as its limits in terms of consistency (a *creature* cannot be *dead* and *alive* at the same time). Divago receives as input at least a pair of concepts and proposes blends of them into new ones (called *blends*). Since the space of possible blends is extremely large, it uses a genetic algorithm to do a parallel search. As a result of such high complexity, it is common that Divago proposes different blends in different runs (of course, eventually repeating itself, depending on the actual complexity of the situation), thus making it very promising in terms of creativity. Divago was tested with several different situations, each other with its own purposes. Three of them were analysed with Ritchie's criteria: the horse-bird experiments, the noun-noun compounds and the creature generation.

The "horse-bird" experiments were meant to test the behaviour of the system with regards to the optimality constraints [Pereira and Cardoso, 2003]. The goal was to find which kinds of configurations were needed to generate a "pegasus", a horse with wings that flies.

The noun-noun compounds experiment aimed to compare to C^3, a system for the interpretation of noun-noun concept combinations [Costello, 1997]. We applied the same noun database that was used for C^3 in testing its capabilities of creative generation of interpretations [Costello and Keane, 2000]. From this comparison, we concluded that, with a rather small set of frames, Divago achieves at least the exact same results as C^3, but it is important to notice that Divago cannot generate the *relational* kind of interpretation (e.g. a "pet fish" is "the fish that is owned by the pet"), in which C^3 has demonstrated to be competent. In terms of a comparison of the creativity of the two systems, the most one can argue is that Divago is able to demonstrate *at least the same level of creativity* [Pereira, 2005] in the sense that it can not only generate, with the same inputs given to C^3 (in an experiment for assessing its creativity [Costello and Keane, 2000]), the same outputs (or similar with small error), but also produce other outputs that C^3 didn't generate. Here, we stress, the application of Ritchie's criteria to C^3 would provide a more extensive comparison of the two systems.

Figure 1: Examples of horse|dragon, horse|werewolf and werewolf|dragon, resp.

The last experiments to report here regard the application of Divago as a server for objects in a game environment. We tested its ability to generate novel creatures for a game, from a database of three creatures (a horse, a dragon and a werewolf). In figure 1, we show some examples, which were generated by a 3D interpreter that receives the concept maps and builds a 3D image.

Each distinct experiment configuration was run 30 times, each run corresponding to a GA entire *evolutive cycle*. The preferred statistical indicator used was the median, as it is not sensitive to outliers (as opposed to the mean) and it usually represents a specific (the median) concept. In terms of creativity, an attempt to apply Ritchie's criteria was made, although with some aspects to focus beforehand: rather than assessing *value* and *typicality*, the pair *novelty*/*usefulness* was preferred. The former was assumed as opposite to typicality and its measure is simply given by edit distance to what is already known to Divago (at the least, the input concepts in a given generation). In other words, each newly generated blend is evaluated against Divago's KB in terms of the number of changes (cut and paste operations) that need to be made to each already known concept, for it to become equivalent to the blend. In Divago, the inspiring set is assumed to correspond to its KB as there are no specific concepts for which it was modelled or inspired by, and the concepts in KB are actually what the system "knows". This implies that typicality is measured by straight comparison to the inspiring set, raising an observation: from the point of view of a creative system, isn't the inspiring set a typicality cluster (or set of clusters) per se? Unless there are concepts in that set explicitly tagged as *wrong* or *untypical*, aren't they implicitly defining

typicality, at least in systems which integrate their inspiring sets (e.g. case-based systems)? Can the same observation be made about value?

Usefulness of a concept depends on how it can satisfy a goal given to Divago (e.g. "give me something that flies and is a transport means"), this being quantified according to the *Relevance* optimality constraint applied [Pereira and Cardoso, 2003]. A correspondence between usefulness and value is much less free of discussion than that of typicality being opposite to novelty. The principle assumed is that something can only be valued according to a purpose or a goal (even if it is to fit an aesthetical matrix). As the system uses Relevance to direct its search, it can be said that usefulness is not an independent measure. This comes as a consequence of the evaluation methodology applied: as blends are (complex) structures, built to satisfy a goal, it does not make sense to evaluate them with respect to other goals. Furthermore, evaluating the results of Divago based on external queries (as happened with Dupond and Wasp) would be at least as controversial as the methodology applied: what would be evaluated (the images, the graphs, the sentences)? what would be value about (image beauty, graph completeness or consistency, sentence syntax or meaningfulness)? The methodology followed was to define a language for queries, which are given externally to the system. The rate of satisfaction of these queries shows the capability of Divago to produce concepts that are useful (in that particular context). This raises two observations: a set of *canonical* problems of creativity (in this case of concept creation) would naturally bring benchmarks to compare with, overcoming these problems of subjectivity in analysis; in systems that are as generic in purpose as Divago is, will it be possible not to include (implicitly or explicitly) some bias towards these benchmarks? If Divago is left without goals, it tends to generate non-sense or simply reproduce the inputs (depending on the configuration).

Results

With the assumptions just given, the fourteen Ritchie's criteria were applied. From the analysis of the values for α, β and γ made for WASP, it was decided to use 0.50 with no other reason than the intuition from being the center of the scale (and having no other reference to compare in similar systems). It is clear, as said above, that these values will change the results of the criteria, however only context and specificities will suggest the best choice. As these were the first experiments in the context of Blending and with Divago, we decided for a choice that intuitively seems more *neutral*. Table 3 presents the numbers obtained.

These criteria may suggest a few conclusions, yet the reader must understand the range of subjectivity involved. Furthermore, while Divago can be seen indeed as demonstrating some creativity, the criteria seem influenced by the large amount of variability of results (namely w.r.t. novelty) and by a possibly simplistic criterion for usefulness. Thus, we make a few qualitative considerations. The system is better at producing good items than typical ones. This has two straight interpretations: since the *Relevance* measure is counted for in the fitness function, Divago drives itself towards its maxima; the space is extremely large and complex, and Divago has no deliberate instruction to search for typicality, therefore the typicality it achieves is more a side effect of respecting constraints and of coinciding with maxima in *usefulness*. The system clearly gets better values for the *Creatures* experiments in comparison to the others. This can be explained by the following facts: the goal given to Divago for the *Creatures* is simpler than for the other two experiments and there is always a global maximum in the search space; the size of the concept maps in the *Creatures* and *Horse-Bird* experiments is bigger than for the *Noun-Noun* experiments, thus non typicality is easier to achieve (with many more choices, the system only by chance will retrieve copies of the inputs). It is remarkable that, even within the same system, there can be found a high variability of values for the criteria. However in relative terms (and by grouping the results by criteria), one can propose a summary of conclusions: Divago was able to generate medium to low typicality (from 1 and 2), medium to high value (from 3 and 4), very high proportion of value within typicality (criterion 5), medium to very high proportion of value within typicality (from 6, 7 and 8), almost no reinventions (9 and 10), medium to low typicality in non reinventions (11 and 13) and medium to high value in non reinventions (12 and 14).

3.3 Dupond

The System

One problem affecting even the most recent sentence and text generators is the narrow diversity of their discourse. A given input usually triggers the same output, no matter how many times it has been submitted before. This hardly happens with human discourse, for humans tend to avoid boring repetitions, by using strategies such as synonymous words to express a given concept, and figures of speech. These and other strategies are still subject to heavy research, and Dupond was developed as part of a wider study on the use of lexical relations to achieve a more natural discourse.

As described in [Mendes *et al.*, 2004], Dupond is a system whose basic functionalities can be summarised as follows: once given a sentence and a set of configuration options, it parses that sentence, disambiguates the words and replaces some of them by synonyms or hypernyms. The output sentences should be different from the input ones, but, ideally, they keep the original meaning unchanged.

Dupond selects the replacement words by following WordNet [Miller *et al.*, 1990] lexical relations and an algorithm that introduces some randomness in the choices. We don't argue that it is creative, although it was developed and evaluated in the light of the most recent theories of creativity.

Results

In order to assess Dupond's results, the system was used to transpose a set of human-written sentences, and those sentences were used to produce enquiries for several different people to classify [Mendes, 2004]. The basic structure of the enquiries was: a human-written sentence was given, and then three computer-generated versions of it for the user to classify. Each of these sentences was considered to have suffered a different transformation, named as follows, for clarity: **N**-Null transformation, i.e., the original sentence; **H1** - Trans-

Crit.	Meaning	Experiment		
		Horse-Bird	Noun-Noun	Creatures
1	Average typicality	0.443	0.543	0.343
2	Typical results / results	0.273	0.563	0.333
3	Average quality	0.504	0.782	1.000
4	Good results / results	0.636	0.781	1.000
5	Good typical results / results	1.000	0.778	1.000
6	Good atypical results / results	0.364	0.344	0.667
7	Good atypical results / atypical results	0.500	0.786	1.000
8	Good atypical results / good typical results	1.333	0.786	2.000
9	Results in the inspiring set / inspiring set	0.000	0.036	0.000
10	Results / results in the inspiring set	N/A	16.000	N/A
11	Average tipicallity of new results	0.406	0.513	0.308
12	Average quality of new results	0.483	0.831	1.000
13	Typical new results / new results	0.273	0.500	0.333
14	Good new results / results	0.636	0.781	1.000

Table 3: Results of Ritchie's Criteria for Divago's experiments, considering $\alpha = \beta = \gamma = 0.5$, val =*usefulness* and typ=$1 - nov$.

formation in which some words were replaced by their first hypernyms; **L** - Transformation in which some words were replaced by their synonym with less senses; **M**- Transformation in which some words were replaced by their synonym with more senses. From now on, for the sake of clarity, we'll say ***X** sentence*, instead of *sentence which suffered transformation **X***.

The sentences could be classified, compared to the first one, in terms of: Originality (**O**), i.e., if the sentence was more original than the first one; Meaning (**S**), i.e., if the sentence meant the same as the first one; and Understandability (**U**), i.e., if the sentence was more comprehensible than the first one. Possible classifications for each sentence were 1 (not at all), 2 (more or less), and 3 (yes, it is). Detailed analysis of the results can be found in [Mendes, 2004].

Application of the Criteria

As stated in section 2, the application of the criteria is based on the concepts of *typicality* and *value*, as well as the inspiring set. In our approach, we assume that *typicality* is the opposite of *originality*, as assumed for Divago and the concept of *novelty*.

A sentence can be considered of value in one of two situations: either if it keeps the meaning of the original one, or if it is more understandable. Therefore, we applied the criteria twice, in order to get a better insight into the system: once considering $val =$ **S**, and a second time considering $val =$ **U**.

The inspiring set is constituted by the set of sentences fed to the system to generate the ones which were used in the enquiries, plus the ones used to test the system during development. Since we used principally four different sentences to test the system during development, this number needs to be added to the number of input sentences to obtain the exact number of elements in the inspiring set. Table 4 summarises the number of elements in the inspiring set, as well as the number of reinventions by Dupond.

Since there are, up to this moment, no recommendations or guidelines for the thresholds α, β or γ, of the Ritchie's Criteria, we also assumed that $\alpha = \beta = \gamma = 0.5$, although there is no guarantee that these are the most appropriate ones to study the system.

Inspiring set	Reinventions		
	H1	L	M
196	3	3	5

Table 4: Inspiring set and reinventions for each transformation (Dupond).

The Value of S

Considering more valuable the sentences whose meaning was closer to the meaning of the sentences which lead to their generation (i.e., $val =$**S**), the results are as shown in table 5.

Analysing the table we can notice that the system's results are *relatively* original (1 and 2) and valuable (3 and 4). **H1** sentences, though, show the smallest value. These sentences, where Dupond used hypernymy relations to find replacement words, are the product of a conceptual generalisation, where some meaning is lost. These generalised sentences are the least valuable ones if one intends to keep the original meaning unchanged.

Only a small amount of the results is good and atypical, hence the system is not very successful in creating original sentences with high value (6 and 7). Only a small proportion of the sentences were reinventions (9), and the new results show average typicality and value (11 and 12). Another point is that, while **H1** sentences show the poorest results, **L** ones show the best.

The Value of U

Considering $val =$ **U**, i.e., the more understandable sentences are more valuable, the results for criteria 3 to 8 are as shown in table 6. The most noticeable difference is that the value of **U** is much lower than the value of **S**, i.e., computer-paraphrased sentences may keep the meaning of the original ones, but are harder to understand. Another point is that **M**

Crit.	Meaning	Transform		
		H1	L	M
1	Average typicality	0.559	0.490	0.554
2	Typical results / results	0.750	0.475	0.650
3	Average quality	0.295	0.495	0.426
4	Good results / results	0.100	0.500	0.400
5	Good typical results / results	0.100	0.474	0.462
6	Good atypical results / results	0.025	0.300	0.125
7	Good atypical results / atypical results	0.091	0.545	0.333
8	Good atypical results / good typical results	0.333	1.333	0.417
9	Results in the inspiring set / inspiring set	0.015	0.015	0.026
10	Results / results in the inspiring set	65.333	65.333	39.200
11	Average typicality of new results	0.559	0.490	0.554
12	Average quality of new results	0.295	0.495	0.426
13	Typical new results / new results	0.750	0.475	0.650
14	Good new results / results	0.100	0.500	0.400

Table 5: Results of Ritchie's Criteria for Dupond's enquiries, considering $\alpha = \beta = \gamma = 0.5$ and val =**S**.

sentences show the higher values, while it was the **L** ones performing better when val = **S**. This means that the sentences using words with more senses are more widely understood, while the ones using words with less senses perform better at keeping the meaning.

Since there're few sentences with a considerable value, then criteria 5 to 8 show residual values, which hardly mean anything important.

Crit.	Transform		
	H1	L	M
3	0.146	0.238	0.294
4	0.000	0.050	0.025
5	0.000	0.053	0.038
6	0.000	0.025	0.000
7	0.000	0.045	0.000
8	∞	1.000	0.000

Table 6: Ritchie's Criteria for Dupond, considering $\alpha = \beta = \gamma = 0.5$, and val = **U**.

4 Discussion

A comparison of the creativity of the systems with these criteria (and possibly any other) has little consistency in the sense that we're dealing with substantially different applications, even though there are many commonalities. One can reach generic conclusions such as "Divago tends to be more creative than Dupond in a variety of criteria," yet one cannot argue that the same conclusion remains valid if they were applied exactly to the same application with same input and configuration. At most, one can aim to propose that following one paradigm (e.g. GAs) rather than the other (e.g. rules), may lead to better results in terms of creativity.

It seems more important to discuss the main problems raised in the application of the criteria. The first one has to do with the rating schemes *val* and *typ*; namely the former would demand a compromise that is rarely explicitly made in everyday observation of creativity, of what a *valuable* outcome is exactly about. This need was already raised (and partially answered) by [Pease *et al.*, 2001], in their proposal of novelty and value ratings, to which some of the experiments presented here agree. Yet again, the main problem is the lack of benchmarks and of standard methodologies. However, typicality can also be ambiguous. In WASP, having the *correct* syntax and number of lines would determine its typicality as a 16th century Spanish poem, while in Divago and Dupond it was assumed that something that is original (i.e. distant from all other items in the KB, in Divago) is not typical. While the analysis of WASP follows the indications of Ritchie more strictly, it becomes simplistic to reduce typicality to syntax and structure. A poem can be untypically Spanish for its content rather than structure. On the other hand, Divago risks calling typical those items that, although structurally incorrect or inconsistent, show no particular originality (i.e. they share many features with a concept from the KB).

There is another problem regarding the variables involved (α, β, γ and θ). We emphasized in table 2 how these values can affect criteria results. Finding acceptable values will depend on experimentation in different contexts. Yet, until now, there has been no application of these. Furthermore, their scales will differ among criteria (e.g. criteria 4 through to 9 yield values in the interval [0, 1], criterion 8 can give any positive real number, criterion 10 always results in values higher than 1). In Divago and Dupond, we assumed α, β and γ to be 0.5 while for WASP, 0.7 0.5 and 0.3 were tested. Another issue (particularly raised in Divago and Dupond experiments) is that Ritchie considers typicality and value, rather than novelty and usefulness. While usefulness and value are often meant as synonymous (in the sense that something is valued when it accomplishes a purpose), typicality runs opposite to novelty (as was assumed in Dupond and Divago).

We also emphasize the inability of the criteria to cope with the iterative nature of computer programs. More precisely, these criteria seem designed for a static evaluation: pick the set of results from one run, then apply the criteria. Following the assumption that the same reasoning could be applied

to a set of runs [Ritchie, 2001], we did so for our systems. However, the problem remains: in the first runs, a system can generate highly valued and atypical results (thus achieving high values in some criteria), then start to repeat itself. Should the first (and highly creative) items made by the system itself be considered as members of the Inspiring Set? Was it creative in the first iterations, and then not creative in the subsequent ones? Such distinctions are hidden in the criteria here discussed. Only criteria that could consider the behaviour throughout a sequence of iterations would solve this issue.

It is fair to say, though, that, in general, if we see the criteria as a set of directives for observing creative products, they represent consensual perspectives within the Computational Creativity field. It is their instantiation that poses the major problems. It is our intuition that the first obstacle for creativity assessment lies not on the criteria, but on the lack of canonic problems to solve. The area is by itself undefined. A set of "creativity challenges" should exist beforehand. If that happened, then possible benchmarks would emerge naturally (and these criteria would be candidates for it).

5 Conclusions

The problem of assessing creative systems is currently in need of research, a consistent proposal being the Ritchie's Criteria—a formalisation of some principles generally accepted in the field of CC, that intend to provide a tool to assess computer programs.

We applied the Criteria to three different systems: a poem generator, a conceptual blender and a sentence paraphraser. The main difficulties noticed were: the concept of *typicality* seemed somehow troublesome—*novelty* or *originality* seem more adequate; troublesome is also the notion of *value*, as it seems to imply a compromise that is not free if controversy—either it becomes reduced to some kind of metric or it becomes a subjective definition: Divago applied a metric, while WASP and Dupond relied on human evaluation; Criteria rely on unspecified parameters (α, β, γ and θ), which cause some difficulties in handling the process and drawing definite conclusions. Another point is that the criteria are designed to assess one time run results, and not to deal with the iterative nature of computer programs.

The lack of canonical problems in the area of creativity, i.e., a set of *typical* problems that would be tackled according to different approaches, emerges also as a conclusion. Indeed, benchmarks and system comparisons can only make sense in highly controlled settings, at least in terms of input data and configurations. The relevance of criteria such as Ritchie's is that, if such problems existed, then benchmarks would naturally emerge.

By divulging these experiments, we hope to contribute to the area with a set of situations to be compared by others in the future, thus providing additional working material and uncovering some practical problems of the proposal.

References

[Colton *et al.*, 2001] Simon Colton, Alison Pease, and Graeme Ritchie. The effect of input knowledge on creativity. In Amílcar Cardoso, Carlos Bento, and Geraint Wiggins, editors, *Proceedings of the First Workshop on Creative Systems, ICCBR*. ICCBR-01, 2001.

[Costello and Keane, 2000] Fintan J. Costello and Mark T. Keane. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349, 2000.

[Costello, 1997] Fintan J. Costello. *Noun-noun conceptual combination: the polysemy of compound phrases*. PhD thesis: Trinity College, Dublin, 1997.

[Fauconnier and Turner, 1998] Gilles Fauconnier and Mark Turner. Conceptual integration networks. *Cognitive Science*, 22(2):133–187, 1998.

[Gervás, 2000] Pablo Gervás. Wasp: Evaluation of different strategies for the automatic generation of Spanish verse. In *AISB-00 Symposium on Artificial Intelligence and Creativite & Cultural Aspects of AI*, 2000.

[Gervás, 2002] Pablo Gervás. Exploring quantitative evaluations of the creativity of automatic poets. In *Proceedings to the Second Workshop on Creative Systems*. European Conference on Artificial Intelligence, ECAI'02, 2002.

[Mendes *et al.*, 2004] Mateus Mendes, Francisco C. Pereira, and Amílcar Cardoso. Creativity in natural language: Studying lexical relations. In *Workshop on Language Resources for Linguistic Creativity*, Lisbon, May 2004. 4th International Conference on Language Resources and Evaluation (LREC).

[Mendes, 2004] Mateus Mendes. Relações lexicais na geração de língua natural. Master's thesis, Dpt. de Engenharia Informática da FCTUC, Universidade de Coimbra, Pinhal de Marrocos, 3000 Coimbra, Portugal, September 2004.

[Miller *et al.*, 1990] George A. Miller, Richard Backwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, pages 234–244, 1990.

[Pease *et al.*, 2001] Alison Pease, Daniel Winterstein, and Simon Colton. Evaluating machine creativity. In Amílcar Cardoso, Carlos Bento, and Geraint Wiggins, editors, *Proceedings of the First Workshop on Creative Systems, ICCBR*. ICCBR-01, 2001.

[Pereira and Cardoso, 2003] Francisco C. Pereira and Amílcar Cardoso. Optimality principles for conceptual blending: A first computational approach. *AISB Journal*, 1(4), 2003.

[Pereira, 2005] Francisco C. Pereira. *A Computational Model of Creativity*. Universidade de Coimbra, 2005.

[Ritchie, 2001] Graeme Ritchie. Assessing creativity. In G. Wiggins, editor, *Proceedings of the AISB'01 Symposiym on Artificial Intelligence and Creativity in Arts and Science*, pages 3–11, 2001.