Defining Creativity: Finding Keywords for Creativity Using Corpus Linguistics Techniques

Anna Jordanous

Creative Systems Lab / Music Informatics Research Centre,
School of Informatics, University of Sussex, UK
a.k.jordanous at sussex.ac.uk

Abstract. A computational system that evaluates creativity needs guidance on what creativity actually is. It is by no means straightforward to provide a computer with a formal definition of creativity; no such definition yet exists and viewpoints in creativity literature vary as to what the key components of creativity are considered to be. This work combines several viewpoints for a more general consensus of how we define creativity, using a corpus linguistics approach. 30 academic papers from various academic disciplines were analysed to extract the most frequently used words and their frequencies in the papers. This data was statistically compared with general word usage in written English. The results form a list of words that are significantly more likely to appear when talking about creativity in academic texts. Such words can be considered keywords for creativity, guiding us in uncovering key sub-components of creativity which can be used for computational assessment of creativity.

1 Introduction

How can a computational system perform autonomous evaluation of creativity? A seemingly simple way is to give the system a definition of creativity which it can use to test whether creativity is present, and to what extent [1, 9, 11].

There have been many attempts to capture the nature of creativity in words [Appendix A lists 30 such papers], but there is currently no accepted consensus and many viewpoints exist which may prioritise different aspects of creativity (this is discussed further in Section 2.1).

Identifying what contributes to our intuitive understanding of creativity can guide us towards a more formal definition of the general concept of creativity. If a word is used significantly more often than expected to discuss creativity, then I suggest it is associated with the meaning of creativity. Many such words may be more tightly defined than creativity itself; we can encode these definitions in a computational test(s) and combine these tests to approximate a measurement of creativity.

The intention of this approach is to make the goal of automated creativity assessment more manageable by reducing creativity to a set of more tractable sub-components, each of which is considered a key contributory factor towards creativity, recognised across a combination of different viewpoints.

2 Finding Keywords For Creativity

The aim of this work is to find words which are significantly more likely to be used in discussions of creativity across several disciplines. These words can be treated as **keywords** that highlight key components of creativity.

What discussions of creativity should be examined? Written text is simpler to analyse than speech and there are many sources to choose from. The texts should be of a reasonable length, otherwise they provide only an overview rather than investigating more subtle points which may be significant. This study concentrates on the academic literature discussing creativity, in order to reduce variability in formats, facilitate discovery of key documents for inclusion and allow a measure of the influence of the document (the number of citations).

To find words used specifically in creativity literature, the language used in several papers was analysed to extract the frequencies with which individual words were used. These extracted word frequencies were statistically compared with data on how the English language is used in general written form.

2.1 Creativity Corpus: A Selection of Papers on Creativity

The academic literature on the nature of creativity ranges over at least the past 60 years; arguably starting from Guilford's seminal 1950 presentation on what creativity is and how to detect it. Many repeated themes have emerged in the literature as important components of creativity. As an example, the word clouds[1] in Figs. 1 and 2 show that the word *new* is frequently used in definitions of creativity and also in discussions of what creativity is.

Fig. 1. Most frequent words in 23 creativity definitions (excluding common-use words)

Wide variance can be found, though, in what are considered primary contributory factors of creativity. For example psychometric tests for creativity (such as [12]) focus on *problem solving* and *divergent thinking*, rewarding the ability to move away from standard solutions to a problem. In contrast, much recent writing in computational creativity (such as [9, 11]) places emphasis on *novelty* and

[1] Generated using software at http://www.wordle.net

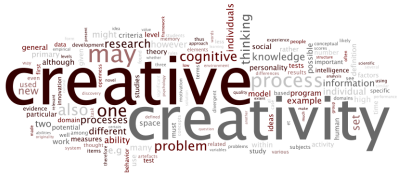

Fig. 2. Most frequently used words in 30 academic papers on creativity (excluding common English words).

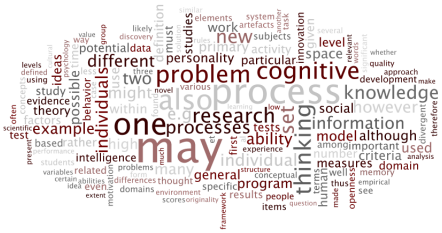

Fig. 3. With *creativity* and *creative* removed (as they dominate the image)

value as key attributes. Whilst there is some crossover, the differing emphases give a subtly different interpretation of creativity across academic fields.

This study considers 30 papers on the nature of creativity, written from a number of different perspectives. This set of papers is referred to in this paper as the *creativity corpus*[2] and is detailed in Appendix A. The 30 papers were selected using criteria such as the paper's influence over future work (particularly measured by number of citations), the year of publication, academic discipline and author(s). To match the diversity of opinions in creativity literature as closely as possible, the set of papers give viewpoints from many different authors, from psychology to computer science backgrounds and across time, from 1950 to the current year (2009). Figure 4 shows the distribution of papers by subject, according to journal classification in the academic database *Scopus*[3].

Years	*num papers*	Psychology	CompSci	Arts&Hums	Other/undefined
1950 – 1969	*3*	3	0	2	0
1970 – 1979	*3*	3	0	1	2
1980 – 1989	*7*	6	0	2	4
1990 – 1999	*6*	4	1	0	3
2000 – 2004	*5*	3	0	0	2
2005 – 2009	*6*	1	4	0	3
	30	20	5	5	14

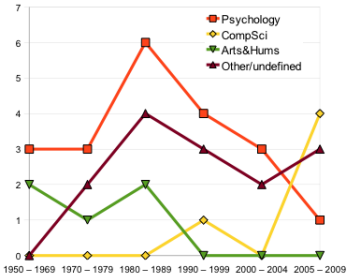

Fig. 4. Distribution of subject area of papers over time

The methodology for this study placed some limitations on what papers could be used. Papers had to be written in English[4] and had to be available in a format that plain text could be extracted from (this excluded books or book chapters).

[2] A *corpus* is the set of all related data being analysed (plural: *corpora*).
[3] Scopus classifies some journals under more than one subject area
[4] All non-British word spellings were amended to British spellings before analysis

2.2 Data Preparation

For each paper a plain text file was generated, containing the full text of that paper. All journal headers and copyright notices were removed from each paper, as were the author names and affiliations, list of references and acknowledgements. All files were also checked for any non-ascii characters and anomalies that may have arisen during the creation of the text file.

2.3 Extraction of Word Frequencies from Data

R is a statistical programming environment[5]that is useful for corpus linguistics analysis. Using R, a word frequency table was constructed from the 30 text files containing the creativity corpus. For each word[6] in the text files, the frequency table listed: how many papers that word is used in and the number of times the word is used in the whole creativity corpus (all papers combined).

2.4 Post Processing of Results

To reduce the size of the frequency table and focus on more important words, all *hapaxes* were removed (words which only appear once in the whole creativity corpus). Any strings of numbers returned as words in the frequency table were also removed. To filter out words that were not used by many authors, any words which appear in less than 5 out of 30 papers were also discarded.

2.5 Analysis of Results

It is not enough to consider purely the word frequencies on their own: a distinction is often made in linguistics [3, 6, 10] between very commonly used words (*form* or *closed class* words) and lower frequency words (*content* or *open class* words): when used more often than usual in a text, the open class words usually hold the most interesting or specific content [3]. So for this study the most common words overall are not necessarily the most useful; as the results in Table 1 show, the most frequent words overall are usually those expected to be prolific in any written texts.

Removing stopwords (very commonly used English words such as "the" or "and") is not sufficient for the purposes of this work: this study focusses on those words which are specifically used more often than expected **when discussing creativity**, as opposed to other texts. A method for quantifying this usage is discussed in the remainder of this section.

[5] http://www.r-project.org/

[6] A word is defined as a string of letters delimited by spaces or punctuation. A compound term such as "problem-solving" was divided into "problem" and "solving".

Data on General Language Use: British National Corpus (BNC). The BNC is a collection of texts and transcriptions of speech, from a variety of sources of British English usage. The corpus comprises approximately 100 million words, of which around 89 million words are from written sources and the remainder from transcriptions of speech. This study only uses data on the written sources, excluding all transcriptions of speech, as the creativity corpus is also solely from written sources. The data used in this study was taken from [7]: relative word frequency data from a sample subset of the written part of the BNC. Before using this data, frequencies were extrapolated to estimate absolute values.

Statistical Testing of Word Frequencies. It was expected that there is a relationship between how many times a word is used in the creativity corpus and how many times it is used in general writing: to use statistical terminology, that the two corpora are correlated. As the data in both corpora is ratio-scored (i.e. the data is measured on a quantifiable scale), a Pearson correlation test can be performed on the word frequency counts for each corpus, to test the hypothesis that there is significant positive correlation.

If there is significant evidence of correlation, then the words which do not follow the general trend of correlation are of most interest: specifically the words that are used more frequently in the creativity corpus than would be expected given the frequency with which they appear in the BNC. A common way to measure this is to use the log likelihood ratio statistic G^2[3, 6, 8, 10][7]:

$$G^2 = 2 \sum o_{ij} (\ln o_{ij} - \ln e_{ij}) \tag{1}$$

o_{ij} = actual observed no of occurrences of a word i in corpus j

e_{ij} = expected no of occurrences of a word i in corpus j (see Eqn. 2):

$$e_{ij} = \frac{(o_{ij} + o_{ik}) * total(j)}{(total(j) + total(k))} \tag{2}$$

$total(j)$ = total number of words in corpus j

The G^2 value is a measure of how well data in one corpus fits a model distribution based on both corpora. The higher the G^2 value, the more that word usage deviates from what is expected given this model.

G^2 measures the extent to which a word deviates from the model but does not indicate which corpus it appears more frequently than expected in. Therefore a subset of the results was discarded: only those words which appear more frequently than expected in the creativity corpus were retained.

[7] An alternative to G^2 is the chi-squared test (χ^2): see [3, 5, 6, 8, 10] for discussion of why G^2 is the more appropriate option for very large corpora.

3 Results

3.1 Raw Frequency Counts

As can be seen by Table 1 and as discussed in Section 2.5, most words which appeared very frequently were common English words, not useful for this study.

Table 1. Most frequently used words in the creativity corpus.

Word	Count in corpus	Word	Count in corpus	Word	Count in corpus
of	8052	is	2412	as	1448
and	4988	that	2372	creativity	1433
to	4420	creative	1994	are	1294
in	3939	for	1716	this	1174
a	3647	be	1561	with	1116

Figure 2 shows the results with "common English words" removed (according to http://www.wordle.com); however as discussed in section 2.5, this study's focus is on how words are used in the creativity corpus compared to normal, so removing only wordle.com's stopwords is not sufficient for our purposes.

3.2 Using the BNC data

As expected, the creativity corpus and BNC word frequencies are significantly positively correlated, at a 99% level of confidence ($p<0.01$). Pearson correlation testing returned a value of $+0.716$.

The results of this study returned 781 words which are significantly more likely to appear in creativity literature then in general for written English (at a 99% level of confidence). Table 2 shows the 100 words with the highest G^2 score.

4 Discussion of Findings

This work has generated a list of words which are significantly associated with academic discussions of what creativity is. The list is ordered by how likely these words are to appear in creativity literature, so the higher they are on the list, the more significantly they are associated with such discussions.

While words such as *divergent* and *originality* have appeared high on the list, as expected, some interesting results have emerged which are more surprising at first glance, for example *openness* is 6th and *empirical* is 21st. One notable observation is that *process*, in 9th position with a G^2 value of 1986.72, is a good deal higher than *product*, in 409th place with a G^2 value of 75.3[8]. Although on closer inspection, the word *process* has been used in more different contexts

[8] Both G^2 values are still well above 6.63, the critical value for significance at $p<0.01$

than *product*, there are still surprisingly many discussions about the processes involved in creativity. This result provides intriguing evidence for the product vs. process debate in creativity assessment [1, 9, 11].

Table 2. Top 100 words in creativity corpus, sorted by descending signed G^2

	Word	G^2		Word	G^2		Word	G^2		Word	G^2
1	creative	17925.3	25	associative	1010.7	49	interactions	661.8	76	subjects	485.1
2	creativity	17242.5	27	influences	962.6	52	criterion	649.8	77	retention	481.3
3	cognitive	4367.8	28	primary	909.2	52	validity	649.8	77	dimensions	481.3
4	domain	2731.4	29	conceptual	902.4	52	according	649.8	79	hypotheses	469.3
5	innovation	2454.6	30	instance	890.4	55	measures	647.2	79	innovative	469.3
6	openness	2165.8	31	developmental	878.4	56	tests	643.9	81	ideas	464.7
7	because	2081.6	32	individual	857.4	57	verbal	637.7	82	related	460.9
8	divergent	1997.4	33	problem	855.3	57	investigations	637.7	83	dimension	457.2
9	process	1986.7	34	intrinsic	854.3	59	heuristics	625.7	83	validation	457.2
10	motivation	1865.0	34	artistic	854.3	59	fluency	625.7	83	attributes	457.2
11	domains	1696.6	36	evolutionary	842.3	59	rated	625.7	86	research	455.3
12	found	1684.5	36	correlated	842.3	62	psychologists	601.6	87	iq	445.2
13	abilities	1528.1	38	ability	832.8	62	complexity	601.6	87	artefacts	445.2
14	thinking	1418.5	39	programs	818.2	64	discoveries	589.6	87	combinations	445.2
15	scores	1395.8	40	intelligence	803.2	64	semantic	589.6	87	predictions	445.2
16	solving	1359.7	41	cannot	782.1	66	discovery	580.4	87	heuristic	445.2
17	individuals	1317.0	41	facilitate	782.1	67	schema	577.6	92	factors	444.6
18	personality	1218.5	43	toward	770.1	67	rat	577.6	93	these	439.6
19	scales	1215.3	44	correlation	746.0	69	unconscious	553.5	94	psychology	423.0
20	processes	1214.0	45	basis	734.0	70	probability	529.4	95	barren	421.1
21	empirical	1191.2	46	computational	721.9	71	self	514.5	96	positively	409.1
22	ratings	1143.1	47	extrinsic	709.9	72	knowledge	504.0	96	investigators	409.1
23	correlations	1046.8	47	selective	709.9	73	variables	496.6	96	perceptual	409.1
24	originality	1022.8	49	cognition	661.8	74	primitive	493.3	99	example	408.3
25	traits	1010.7	49	hypothesis	661.8	74	novelty	493.3	100	elements	406.5

Some words appear surprisingly highly in Table 2, due to unexpectedly low frequencies being recorded in the BNC data. Two examples are *because* and *found*. This suggests two possibilities: either a slight weakness in the representativeness of the sample BNC data from [7] (perhaps understandable given the sheer quantity of data in the BNC; no sample can be 100% representative of a larger set of data), or alternatively these words may be used more in academic writing than in everyday speech - see section 4.1 for further discussion of this.

From inspection, such words seem relatively infrequent, however, compared to the large number of words which are recognisably associated with creativity in at least some academic domains.

4.1 Further Exploration of Keywords

Words in Common Academic Usage. It is possible that some words feature highly in the results solely because they are common academic words. Therefore the results list should be compared to common academic words to see if there

is evidence of correlation between the two sets of data. If so, this should also be taken into account.

Two lists of common words in academic English were found: the Academic Word List (AWL) [2] and the University Word List (UWL) [13]. Both contain groups of words, in order of frequency of usage specifically in academic documents (group 1 holds the most frequent words). Unlike the BNC corpus, the AWL and UWL only provides summary information on academic word usage with no actual frequency data per word; this limits what statistical testing can be performed. Spearman correlation testing returns a value of -0.236 correlation between the creativity corpus and the AWL and -0.210 correlation between the creativity corpus and the UWL. Neither correlation value is significant at $p<0.01$ (or $p<0.05$). As this indicates no significant relationship between the creativity corpus and either academic list, no correction should be made to the keyword results on account of either set of academic data.

Poor availability of any other data on academic word usage hinders further investigation of this issue at present.

Context and Semantics. Although the list of keywords hold much of interest in uncovering what is key to creativity, they rely purely on frequency of word usage. The results are not intended to account for the different contexts in which words are used; when analysing large corpora, exploring every word's semantic context would be highly time-consuming. Instead, the frequency results highlight keywords to focus on in the texts and examine in more detail [6, 10].

Categorising the keywords by semantics is non-trivial and "labour-intensive" [4]. Carrying this out empirically would be a significant step in itself and is a fruitful avenue for further work. From inspection of the contexts in which keywords are used, some key categories are suggested in Table 3.

5 Conclusions

For a computational system to be able to perform automated assessment of creativity by a computational system, it needs some point of reference on what creativity is. There is no accepted consensus on the exact definition of creativity. This work empirically derives a set of keywords that combine a variety of viewpoints from different perspectives, for a more universal encapsulation of creativity.

Keywords were calculated through corpus analysis of 30 academic papers on the nature of creativity. The likelihood measure G^2 (Eqn. 1) was used to compare word frequencies in the creativity papers against usage of those words in general written English, as represented by the sub-corpus of the BNC containing written texts (see Section 2.5). This analysis returned 781 words which were statistically more common in the creativity literature sample than expected, given their general usage in written English. Table 2 displays the top 100 results.

The list of keywords encapsulates words we commonly use to describe and analyse creativity in academia. Given their strong association with creativity,

Table 3. Key categories for creativity, generated through examining the keywords

Category	Keywords representing this category
cognitive processes	thinking, primary, conceptual, cognition, perceptual
originality	innovation, originality, novelty
the creative individual	personality, motivation, traits, individual, intrinsic, self
ability	solving, intelligence, facilitate, fluency, knowledge, IQ
influences	influences, problem, extrinsic, example, interactions, domain
divergence	divergent, investigations, fluency, ideas, research, discovery
autonomy	unconscious
discovery	openness, awareness, search, discovery, fluency, research
dimensions	dimensions, attributes, factors, criterion
association	associative, correlation, related, combinations, semantic
product	artefacts, artistic, elements, verbal
value	motivation, artistic, solving, positive, validation, retention
study of creativity	empirical, predictions, tests, hypothesis, validation, research
measuress of creativity	scores, scales, empirical, ratings, criterion, measures, tests
evolution of creativity	developmental, primary, evolutionary, primitive, basis
replicating creativity	programs, computational, process, heuristics

they point us towards sub-components of creativity that contribute to our intuitive understanding of what creativity is.

Many of the keywords in the results can be tested for by a computer more easily than testing for creativity itself. For example:

- Originality: Comparing products to other examples in that domain or to a prototype, to measure similarity
- Ability: Depending on the domain, there are usually many standardised tests to measure competence in that domain
- Divergence: Measuring variance of products against each other
- Autonomy: Quantifiying the assistance needed during the creative process
- Value: Again this is domain dependent and there will usually be many tests for value measurement in a particular domain

The results presented in this paper identify key components of creativity through a combination of several viewpoints. These rsesults will be used to guide experiments implementing a computational system that evaluates creativity by testing for the key categories that have been identified. The experiments enable us to determine whether this approach to defining creativity gives a good enough approximation for creativity evaluation, and if so, which combination of tests most closely replicates human assessment of creativity.

Acknowledgements

Nick Collins, Sandra Deshors, Clare Jonas and Luisa Natali all made useful comments during discussions of this work.

References

[1] S. Colton. Creativity versus the perception of creativity in computational systems. In *Proc. of AAAI Symposium on Creative Systems*, pages 14–20, 2008.

[2] A. Coxhead. A new academic word list. *TESOL quarterly*, 34(2):213–238, 2000.

[3] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[4] D. Glynn. Polysemy, syntax, and variation. In V. Evans and S. Pourcel, editors, *New Directions in Cognitive Linguistics*. John Benjamins, Amsterdam, 2009.

[5] S. T. Gries. Null-hypothesis significance testing of word frequencies: a follow-up on kilgarriff. *Corpus Linguistics and Linguistic Theory*, 1(2):277–294, 2005.

[6] A. Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.

[7] G. Leech, P. Rayson, and A. Wilson. *Word Frequencies in Written and Spoken English*. Longman, London, UK, 2001.

[8] M. P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, UK, 1998.

[9] A. Pease, D. Winterstein, and S. Colton. Evaluating machine creativity. In *Proc. of ICCBR Workshop on Approaches to Creativity*, 2001.

[10] P. Rayson and R. Garside. Comparing corpora using frequency profiling. *Proc. of ACL Workshop on Comparing Corpora*, 2000.

[11] G. Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99, 2007.

[12] E. P. Torrance. *Torrance tests of creative thinking*. Scholastic Testing Service, Bensenville, IL, 1974.

[13] G. Xue and I. S. P. Nation. A university word list. *Language Learning and Communication*, 3(2):215–229, 1984.

Appendix A: Papers in the Creativity Corpus

Year	Paper title	Author(s)	Reference
1950	Creativity	Guilford	*Am Psychol 5 (9), 444-454*
1960	Blind variation and selective retentions in creative thought as in other knowledge processes	Campbell	*Psychol Rev 67 (6), 380-400*
1962	The associative basis of the creative process	Mednick	*Psychol Rev 69 (3), 220-232*
1970	Identification of creativity: The individual	Dellas and Gaier	*Psychol Bull 73 (1), 55-73*
1970	Identification of potentially creative persons from the Adjective Check List	Domino	*J Consult Clin Psychol 35, 48-51*
1979	A creative personality scale for the Adjective Check List	Gough	*J Pers Soc Psychol 37 (8), 1398-1405*
1980	Primary process thinking and creativity	Suler	*Psychol Bull 88 (1), 144-165*
1983	The social psychology of creativity: A componential conceptualization	Amabile	*J Pers Soc Psychol 45 (2), 357-376*
1987	Creativity, divergent thinking, and openness to experience	McCrae and Ingraham	*J Pers Soc Psychol 52 (6), 1258-1265*
1988	Assessing everyday creativity: Characteristics of the Lifetime Creativity Scales and validation with three large samples	Richards, Kinney, Benet & Merzel	*J Pers Soc Psychol 54 (3), 476-485*
1988	Creativity syndrome: Integration, application, and innovation	Mumford and Gustafson	*Psychol Bull 103 (1), 27-43*
1988	Motivation and creativity: Toward a synthesis of structural and energistic approaches to cognition	Csikszentmihalyi	*New Ideas in Psychol 6 (2), 159-176*
1988	The creative mind: Toward an evolutionary theory of discovery and innovation	Findlay and Lumsden	*J Soc Biol Struct 11 (1), 3-55*
1992	The psychoeconomic approach to creativity	Rubenson and Runco	*New Ideas in Psychol 10 (2), 131-147*
1994	Precis of The Creative Mind	Boden	*Behav Brain Sci 17 (3), 519-570*
1995	Cognition and creativity	Runco and Chand	*Educ Psychol Rev 7 (3), 243-267*
1996	A theory of individual creative action in multiple social domains	Ford	*Acad Manage Rev 21 (4), 1112-1142*
1996	Creativity and the five-factor model	King, Walker and Broyles	*J Res Pers 30 (2), 189-203*
1996	Creativity, emergence and evolution in design	Gero	*Knowl Based Syst 9 (7), 435-448*
2000	Creativity: Cognitive, personal, developmental, and social aspects	Simonton	*Am Psychol 55 (1), 151-158*
2001	The effect of input knowledge on creativity	Colton, Pease and Ritchie	*Proc ICCBR workshop (Creative Sys)*
2002	Aspects of a cognitive theory of creativity in musical composition	Pearce and Wiggins	*Proc ECAI workshop (Creative Sys)*
2004	The cognitive neuroscience of creativity	Dietrich	*Psychon Bull Rev 11 (6), 1011-1026*
2004	Why isn't creativity more important to educational psychologists	Plucker, Beghetto and Dow	*Educ Psychol 39 (2), 83-96*
2006	A preliminary framework for description, analysis and comparison	Wiggins	*Knowl Based Syst 19 (7), 449-458*
2006	Can we trust creativity tests: a review of the Torrance Tests of Creative Thinking	Kim	*Creativity Res J 18 (1), 3-14*
2006	The transformational creativity hypothesis	Ritchie	*New Gen Com 24 (3), 241-266*
2007	Intuition, insight, imagination and creativity	Duch	*IEEE Comput Intell Mag 2 (3), 40-52*
2007	Some empirical criteria for attributing creativity to a computer program	Ritchie	*Minds Mach 17 (1), 67-99*
2009	Creativity map: Toward the next generation of theories of creativity	Ivcevic	*Psychol Aesth Creat Arts 3(1), 17-21*