Text Analysis of Resumes and Lexical Choice as an Indicator of Creativity

Alexander Rybalov

Infibond Corp.,8 HaBarzel Street, Tel Viv, Israel

Abstract

Contemporary human resources departments use text analysis of resumes to determine whether job applicants are suited for certain positions. We computed a similarity measurement function to compare resumes of applicants of varying education levels in different job categories. A high correlation was found with trust, cognition and positive emotions. We showed that standardized templates are often designed by applicants to elicit positive emotions about their job candidacy and to suggest greater trust and cognitive abilities. Nevertheless, these measurements have a negative correlation with education level. The implication is that applicants that use standardized templates for writing resumes have lower levels of creativity.

1. Introduction

Applicants oftentimes write resumes using standardized templates (in the first 50 references in Google search on 'patterns' and 'resumes,' the vast majority offer templates for resume writing). This raises two questions: 1) What features are they trying to convey to potential employers when they use these templates and how to quantify these connections? 2) Does the use of standardized patterns indicate lower creativity?

To answer the first question – what features they trying to convey to potential employers when they use these templates and how to quantify these connection – the model that links similarities between resumes with personal features derived from text analysis. The most salient feature is trust [1, 2]. Another feature is cognition – recruiters are more likely to accept applicants with high cognitive abilities [3]. For these two features we can compute how semantic distance between resumes and words related to these features. Words describing cognitive process are taken from LIWC (Linguistic Inquire and Word Count) [4], whereas words related to trust are from [5]. Another group of parameters related to cognition are syntactic features, in particular prepositions and conjunctions [6, 7]. Since impressions made by resumes influence hiring decisions, the third group of parameters is based on sentiments analysis [8].

The answer to the second question – does the use of standardized patterns indicate lower creativity – is linked to distinguishing between domain-related knowledge, cognitive skills and creativity-relevant skills. Domain-related knowledge represent factual knowledge, technical skills, and special talents in the problem area in question [9]. It includes facts, principles, attitudes toward various issues in the domain, knowledge of paradigms, techniques, and methods for solving problems in the domain, and aesthetic criteria. "This component can be viewed as the set of cognitive pathways for solving a given problem or doing a given task". Such pathways or algorithms represent set sequences of individual steps for performing tasks or solving problems in a given domain. Therefore, domain-relevant cognitive skills facilitate solving of problems within particular contexts or domains and can facilitate habitual problem-solving even when the problem at hand is novel [10].

Domain-relevant cognitive skills depend on a number of personal qualities such as innate cognitive, perceptual, and motor abilities as well as the formal and informal education of the individual in a specific domain. “Creativity-relevant skills include a cognitive style favorable to taking new perspectives on problems, application of heuristics for the exploration of new cognitive pathways, and a working style conducive to the persistent, energetic pursuit of one’s work”. In other words, individual creativity is facilitated by the ability of an individual to create unusual interpretations of the issues at hand [10]. These skills build the “something extra” of creative performance. And vice-versa, people with lower creativity are likely to work on tasks that are well-defined as opposed to ill-defined tasks. Since using standardize templates for writing resumes is well- defined task, the greater utilization of them implies lower creativity.

2. Related work

There are a number of works that discuss influence of trust on human behavior, in particular participation in social networks. The paper of C.-N. Ziegler and J. Golbeck [1] analyzes trust relationship and how they related to user similarity. The authors define trust through social networks and discuss how it influences recommendation systems. By analyzing a movie rating website, they conclude that as trust among the site’s users increases, disparities in the scores they assign given films narrow, demonstrating a strong correlation between trust and user similarity. This has implications on the resume process, as being able to convey trust through the language of a screened CV might attract more hirers in HR to advance applicants in the job process.

The paper [2] discusses how trust influences human behavior in general, and in social networks specifically. The authors define trust as a mediated value of opinions belonging to someone’s acquaintances. When designing models of relationships between two people, the authors elect to refer to the collected information about one of the parties as a “resume,” namely because it represents similar biodata such as degrees, work history, personal skills, and attitudes among other personal features, just as in a professional resume. The authors then introduce relevance and correspondence factors that take into account the similarity of resumes. Specifically, it evaluates a person’s relating the resume of a potential acquaintance with that of current friends and peers in order to assess trustworthiness of the candidate acquaintance.

In [11] the analysis confirms that education is strongly related to trust. However, most of this connection is explained by cognitive ability and the resulting occupational prestige linked to the level of educational attainment.

One of the few papers that where text analysis was applied to analysis of resumes is the dissertation of Joshua D. Weaver [6] published in April 2017. In his paper, job performance was predicted on a textual analysis of resumes. However, certain issues in the data raise doubts about the conclusions of the paper. For instance. the average task performance of blue-collar workers was much higher than so-called pink-collar (service industry) workers, which in turn was higher than the task performance of white-collar workers. As a result, the value of R^2 in his prediction model is equal to 0.065. In addition, the value of variables (differentiation words, conjunctions, words longer than six characters, prepositions, cognitive process words) that are used to build a composite score of Written Cognitive Ability Index (WCAI) is much higher for applicants with trade, vocational, technical school or just high school education than for professionals or associates with academic post-high school education. Thus, WCAI predicts

that applicants with less education and experience will achieve superior task performance. This paper, where analysis was done across all resumes, has already elicited references in other academic paper, where analysis was done across all resumes, has already elicited references in other academic paper, where analysis was done across all resumes, has already elicited references in other academic papers [12, 13]. Even though it is pioneering work that applies natural language processing to the analysis of resumes, up to our knowledge its conclusions are not correct and the research should be incorporated carefully when reviewed ahead of future studies.

Differentiation of cognitive skills into domain-related knowledge and creative thinking abilities was first introduced in [14]. Later on this approach was extended in [15], where the authors try to identify contextual and personal factors that may foster or hinder creativity. Employee creativity has been seen as a critical impetus for organizational innovation.

In the paper [16], the authors demonstrate that solving well-defined problems reduces performance of subsequent creative tasks on "ill-defined problems." The research assesses the impact of working on well-defined or guided tasks prior to trying to work on more abstract or ill-defined tasks, as well as the preference of subjects when choosing between these two varieties of tasks to perform next. This factor (preference for well-defined problems) reduces divergent thinking ability. The presence of defined goal rather than a known set of operators is to blame for this decline. Self-perceived creativity determines whether a person will select a well-structured task as opposed to an ill-structured task. Thus, people that prefer to follow well-structured tasks (e.g. compiling resumes using templates) are likely to work with well-defined problems, and consequently exhibit lower creativity. Cognitive processes activated to find solutions to well-structured problems remain 'in effect' when performing subsequent tasks, thus inhibiting creativity needed for resolving less well-defined tasks.

3. Methods

We use 1,276 resumes of applicants with different levels of education, that after cleaning and excluding duplicates ae reduced to 905 (417 with no college education, 407 bachelor's degrees, 166 master's degrees, and 61 PhDs) and across a number of professions.

To measure the extent to which applicants use templates, the similarity of resumes was calculated. Similarity between resumes was calculated using standard techniques of similarity, using the total frequency – inverse document frequency (TF-IDF) measure.

Similarity measurement [17] is defined as cosine similarity between TF-IDF (total frequency – inverse document frequency) of each pair of resumes, and then averaging across all resumes: *TF-IDF* measures how important a given word is in a resume.

TF-IDF is calculated as

$$TFIDF = TF \cdot IDF$$

Where TF is *total frequency* equal to the number of times that given word happened in resume and IDF is *inverse document frequency* equal to the ratio of the total number of resumes to the number of resumes containing that word, and then taking the logarithm of that ratio.

Once we calculate TF-IDF of each word *k*, we can compute the similarity of two given resumes res, defined as cosine similarity:

$$sim(res^i, res^j) = \frac{\sum_{k=1}^{n} TFIDF(res_k^i) \cdot TFIDF(res_k^j)}{\sqrt{\sum_{k=1}^{n} [TFIDF(res_k^i)]^2} \cdot \sqrt{\sum_{k=1}^{n} [TFIDF(res_k^j)]^2}}$$

where $TFIDF(res_k^i)$ is TF-IDF of the word k in resume i, $k = 1, \dots n$, n is the number of unique words.

For each applicant, the average of similarities to all other resumes was computed. The resulting value was taken as a measure of the extent to which this applicant used standardized templates. Similarity(res^i) of each resume res^i is defined as the average of all similarities of this resume:

$$similarity(res^i) = \frac{1}{m} \sum_{j=1}^{m} sim(res^i, res^j)$$

where m is the number of resumes.

Resume similarity is equal to mean of pairwise distances between a particular resume and all resumes and, intuitively, measures to which degree an applicant uses standard template to write his/her resume.

Independent variables form three groups.

The first group include number of prepositions and conjunctions per length of a text. We evaluate known approach: a number of prepositions and conjunctions as prediction of cognitive ability [6, 7] or task performance.

The second group comprise two variables related to cognitive ability were semantic similarity (semantic distance) between resumes and words associated with trust and cognitive processes. We use semantic distances between text of resumes and sets of words that are highly correlated with cognitive process and trust , were used as proxies for cognition and trust. Another group of features was derived from sentiment analysis of resumes and from linguistic analysis.
Word associated with cognitive processes were taken from Linguistic Inquiry and Word Count (LIWC) [4]. Words associated with trust were taken from [5].

Semantic distance was calculated as sentence similarity [18]. The latter uses semantic similarity between words. Wordsare organized into synonym sets (synsets) in the knowledge base in WorldNet [19].

These synsets form the hierarchical semantic network. Similarity between words is calculated as the minimum length of a path connecting the two words multiplied by depth in the semantic hierarchy of the most specific concept which is an ancestor of both words. Semantic similarity between texts is based on similarity between semantic vectors created from these texts. In this study semantic similarity was calculated between resumes and words highly correlated with trust [5]. Both resumes and words highly correlated with trust were structured as bag-of-words.

The third group of independent variables are related to sentiment. Sentiment scoring was done using the VADER package in PYTHON which determines positive, negative, and neutral scores, as well as a compound (aggregated) score. A determination of the extent to which resumes have psychological traits was done using semantic similarity between resumes and words associated with each of the abovementioned traits: trust and cognition.

4. Evaluation

Since applicants that belong to the same category are likely to use similar patterns, we group resumes according to job categories, whose distribution is shown in the table below:

Table 1. Job categories

Job category	Frequency
Information Technology	90
Engineering	82
Education	69
Management	68
Health & Fitness	61
Sales	53
Designing	48
Finance	47
Digital Media	46
Accountant	42
Business Development	36
HR	30
Banking	29
Arts	29
Advocate	28
Automobile	24
Consultant	23
Food & Beverages	20
Agricultural	20
BPO	17
Aviation	11
Building & Construction	10
Public Relations	10
Architects	7
Apparel	5

Thus, once we calculate the similarity of resumes only to resumes belonging to the same category, we can see very interesting patterns:

Table 2. Correlation between similarity of resumes and independent variables by education level

	Part of speech		Sentiment Scores				Semantic distance	
	prepositions	conjunctions	negative	neutral	positive	compound	cognitive	trust
all	0.04	-0.01	-0.07	-0.29	0.33	0.25	0.49	0.55
no college education	0.05	0.03	-0.03	-0.26	0.28	0.26	0.49	0.65
bachelors	0.05	-0.05	-0.07	-0.33	0.37	0.29	0.49	0.47
masters	-0.01	-0.10	-0.11	-0.25	0.31	0.17	0.54	0.57
PhDs	0.20	-0.03	-0.14	-0.10	0.20	0.18	0.48	0.47

As can be seen, standardized templates are designed to bring about positive emotions (correlation of positive emotions with similarity is 0.33) and trust (correlation is equal to 0.55) as well as convey presumed high cognitive abilities (correlation is equal to 0.49).Words associated with positive emotions are put in resumes primarily at the expense of neutral words.

We can check histograms of which variables have normal distribution using Kolmogorov-Smirnov test.

Table 3. Kolmogorov-Smirnov test of similarity and independent variables

variable	statistics	p-value
similarity	0.22	0.02
prepositions	0.20	0.03
conjunctions	0.15	0.19
trust	0.19	0.04
cognitive	0.34	0.00
negative	0.27	0.00
neutral	0.18	0.06
positive	0.14	0.28
compound	0.43	0.00

As can be seen only positive sentiment and conjunctions have normal distribution. However, checking whether histograms of variables have log-normal distribution produces different results.

Table 4. Kolmogorov-Smirnov test whether similarity and independent variables have log-normal distribution

variable	statistics	p-value
similarity	0.18	0.06
prepositions	0.34	0.00
conjunctions	0.13	0.34
trust	0.15	0.17
cognitive	0.34	0.00
negative	0.55	0.00
neutral	0.15	0.17
positive	0.14	0.26
compound	0.34	0.00

Histograms of trust and positive sentiment, as well as conjunctions and neutral sentiment have lognormal distribution. In addition, p-values for lognormal distribution are larger.

Figure 1. Histogram of trust

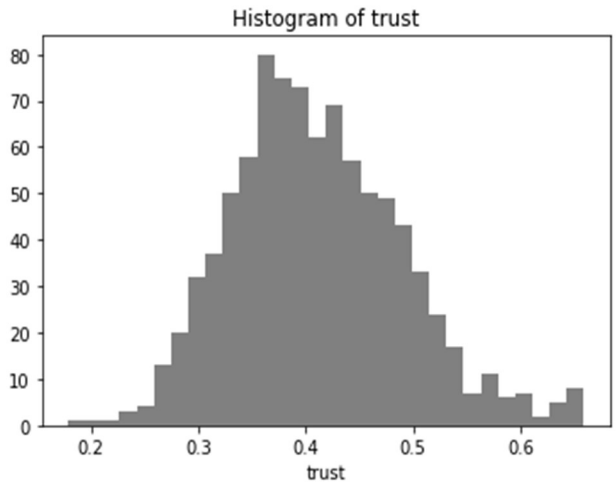
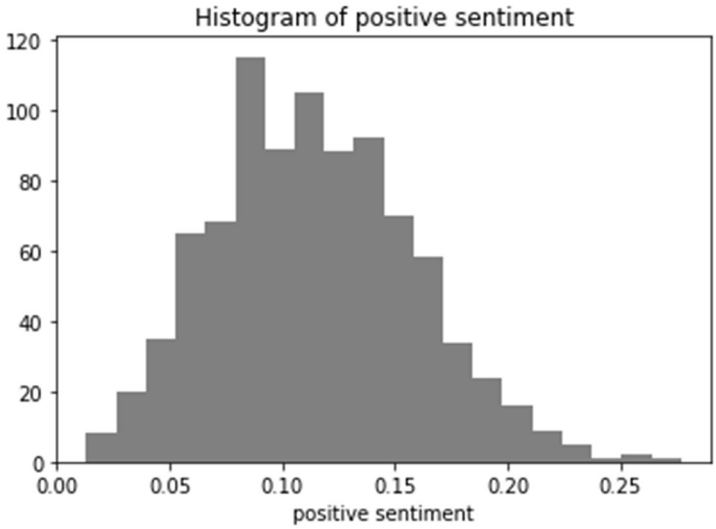


Figure 2. Histogram of positive sentiment



To select independent variables that should be used in model we apply sequential feature selection algorithm that makes evaluation of feature subsets. The best subset was selected by optimizing R^2 (square loss) metric for lasso and ridge regression. The selected variables were 1) positive sentiment and 2) semantic distance with trust.

The chosen model is lasso and ridge regression that achieve 0.40 value of R^2 (square loss) on the test set (one third of total number of resumes). Thus, from textual analysis we can predict the extent to which each resume follows established patterns and draw conclusions about the potential creativity of a given applicant.

Parameters that define similarity are different for applicants with dissimilar levels of education, as can be seen in the following diagrams.

Figure 3. Similarity versus positive sentiment

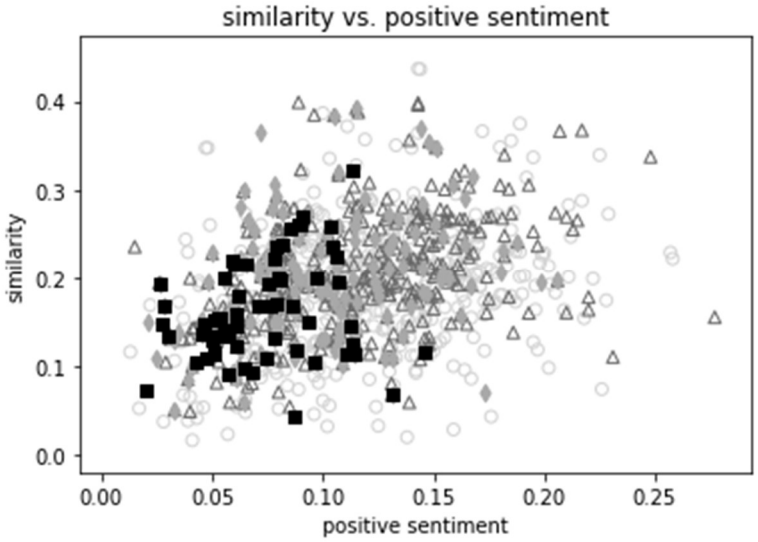
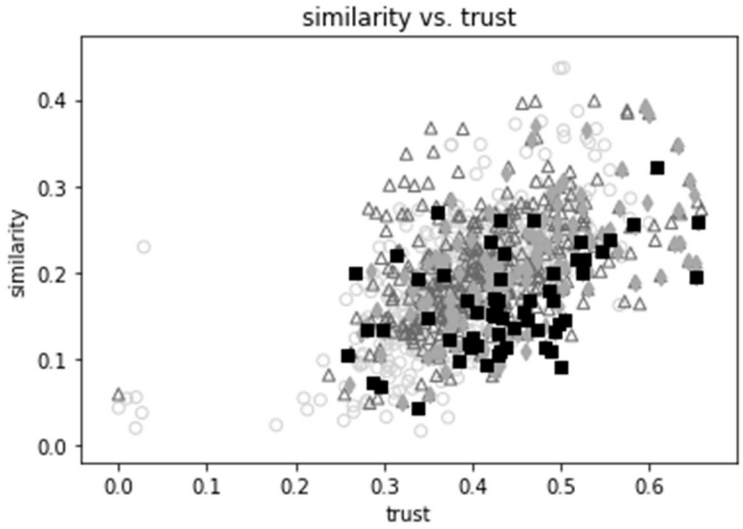


Figure 4. Similarity versus trust



PhDs ■ Masters ▲ Bachelors △ Applicants without college education ○

As can be seen the people with higher education (PhDs and Masters) tend to be more concentrated in lower left part of the diagram, corresponding to lower level of similarity and lower level of positive emotion words in resumes.

Table 5. Mean values of dependent and independent variables versus educational status

	No college education	Bachelors	Masters	PhDs
similarity	0.19	0.2	0.2	0.17
negative	0.02	0.01	0.01	0.02
neutral	0.86	0.87	0.89	0.91
positive	0.12	0.114	0.098	0.075
compound	0.93	0.96	0.95	0.93
cognitive	0.25	0.25	0.26	0.25
trust	0.39	0.42	0.45	0.43

For people with no college education, the average value of positive emotions is 0.120, and for those with a bachelor's degree it is 0.114, whereas for master's degree holders it is equal to 0.098, while for PhDs it is 0.075. Thus, the average value of similarity of resumes for PhDs is equal to 0.16, as opposed to master's (0.196), bachelor's (0.201), and persons that don't have a college degree (0.192).

People with no college education use patterns that try to elicit trust (correlation with similarity is 0.65 – the highest value across all education categories), whereas bachelor's degree earners – besides trust (correlation is 0.48) – try to elicit positive emotions (correlation with similarity is 0.36 – the highest value across all education categories).

For people with no college education, the average value of positive emotions is 0.120, and for those with a bachelor's degree it is 0.114, whereas for master's degree holders it is equal to 0.098, while for PhDs it is 0.076. Thus, the average value of similarity of resumes for PhDs is

equal to 0.16, as opposed to master's (0.196), bachelor's (0.201), and persons that don't have a college degree (0.192).

People with no college education use patterns that try to elicit trust (correlation with similarity is 0.65 – the highest value across all education categories), whereas bachelor's degree earners – besides trust (correlation is 0.48) – try to elicit positive emotions (correlation with similarity is 0.36 – the highest value across all education categories).

5. Conclusions

Standardized templates are designed to bring about positive emotions (correlation of positive emotions with resume similarity is 0.33) and trust (correlation is equal to 0.55) as well as made positive impression about their cognitive abilities (correlation with resume similarity is equal to 0.49). Words associated with positive emotions are put in resumes primarily at the expense of neutral words. For people with no college education, the average value of positive emotions is 0.120, and for those with a bachelor's degree it is 0.114, whereas for master's degree holders it is equal to 0.098, while for PhDs it is 0.076. Thus, the lower level of education of an applicant the more he/she try to elicit positive emotions to increase probability of hiring.

Applicants with no college education use standardized templates that try to elicit trust more than applicants with any level of college education (correlation with resume similarity is 0.65 – the highest value across all education categories), whereas applicants with bachelor's degree – besides trust (correlation with resume similarity is 0.48) – try to elicit positive emotions (correlation with resume similarity is 0.36 – the highest value across all education categories).

Standardized resume templates are more likely to be used by people that, for whatever reason, do not want to invest too much effort in writing resumes. These results confirm that following standardized templates indicates lower creativity. Such patterns attract people with well-defined problem-solving mindsets characterized by convergent thinking, that have single, correct and/or appropriate answer [Downstream]. Their defining feature of convergent thinking that is most effective in solving well-defined problems because it "emphasizes speed, accuracy, logic" in pursuit of "the single best (or correct) answer to a clearly defined question" [20]. Employees that have convergent thinking prefer to engage in routine work, whereas employees who choose ill-defined tasks like to take part in imaginative work. Employees consistently engaged in routine work would produce less-creative ideas than those who were not so engaged. In addition, people are unlikely to switch from one type of work to another [Downstream].

Assumption that preference for well-defined tasks, such as using standardized templates to write resumes is reflection of lower creativity, is supported by the fact that PhDs which have the lowest level of convergent thinking also have lowest use of standardized templates. Skills that require creative thinking include a cognitive style and personality characteristics that are conducive to independence, risk-taking, and taking new perspectives on problems, as well as a disciplined work style and skills in generating ideas. These cognitive processes include the ability to use wide, flexible categories for synthesizing information and the ability to break out of perceptual and performance "scripts" [21].

References

1. Ziegler C.-N. and Golbeck J. : Investigating Correlations of Trust and Interest Similarity – Do Birds of a Feather Really Flock Together?, Decision Support System (2005)
2. Carchiolo V. , Longheu A. , Malgeri M., Mangion G. I, Nicosia V.: The Dilemma of Trust: A Social Network Based Approach, Scalable Computing: Practice and Experience (2008), Volume 9, Number 1, pp. 29–37
3. Cole M.S., Feild H.S., Giles W.F.: Using recruiter assessments of applicants' resume content to predict applicant mental ability and big five personality dimensions, International Journal of Selection, Vol. 11, Number 1, March 2003
4. Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin
5. Yarkoni Tal : Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers, Journal of research in personality, 2010, June 1; 44(3): 363–373
6. Weaver, Joshua D.: Predicting Employee Performance Using Text Data from Resumes, Seattle Pacific University, 2017
7. Zablotskaya Ksenia, Automatic Estimation of Users' Verbal Intelligence, Ulm University, 2015
8. Hisaki Shiraishi: Mutual Relationship of Human Resource Management and Technology, Management and Technological Challenges in the Digital Age, Edited By Pedro Novo Melo, Carolina Machado, Taylor & Francis, 2018
9. Amabile, T. (1988) A model of creativity and innovation in organizations. Research in Organizational Behavior, Vol. 10, 123-167
10. Ford, C. (1996) A theory of individual creative action in multiple social domains. Academy of Management Review, Vol. 21, No. 4, 1112-1142
11. Hooghe, M., Marien, S., De Vroome, T. (2012). The cognitive basis of trust. The relation between education, cognitive ability, and generalized and political trust, Intelligence, 40 (6), 604-613
12. Armstrong Michael Beaumont : Word Counts in Response to Cognitively Demanding Essay Prompts as Reflections of General Cognitive Ability and Broad Cognitive Abilities, Old Dominion University 2018
13. Auer Elena Margaret Lawrence: Detecting Deceptive Impression Management Behaviors in Interviews Using Natural Language Processing, Old Dominion University 2018
14. Amabile, T.: The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology (1983), Vol. 45, No. 2, 357-377
15. Qin Zhou, Shipton Helen: Making Creativity an Attractive Option, Human Resource Management, Innovation and Performance (2016), pp 313-327
16. Moreau C. Page, Engeset Marit Gundersen: The Downstream Consequences of Problem-Solving Mindsets: How Playing with LEGO Influences Creativity, Journal of Marketing Research, Vol. LIII (February 2016), 18-30
17. Elias Iosif, Alexandros Potamianos: Unsupervised Semantic Similarity Computation Between Terms Using Web Documents, IEEE Transactions On Knowledge And Data Engineering, 2010
18. Yuhua Li, McLean David, Bandar Zuhair A., O'Shea James D., Crockett Keeley: Sentence Similarity Based on Semantic Nets and Corpus Statistics, IEEE Transactions on Knowledge & Data Engineering, vol. 18, Issue No. 08 August 2006

19. Miller G.A.: WordNet: A Lexical Database for English, Comm. ACM, vol. 38, no. 11, pp. 39-41, 1995
20. Cropley, Arthur: In Praise of Convergent Thinking, Creativity Research Journal (2006), 18 (3), 391-404
21. Amabile Teresa M., Harvard Business School: Componential Theory Of Creativity, Encyclopedia Of Management Theory (Eric H. Kessler, Ed.), Sage Publications, 2013, 133-139