INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED'07
28 - 31 AUGUST 2007, CITE DES SCIENCES ET DE L'INDUSTRIE, PARIS, FRANCE

A TEXT-BASED ANALYSIS APPROACH TO REPRESENTING THE DESIGN SELECTION PROCESS

Maria C. Yang[1] and Haifeng Ji[2]
[1]Massachusetts Institute of Technology
[2]University of Southern California

ABSTRACT
The language that designers use reflects the design process that they engage in, and analysis of this language can provide valuable insights into the design process. The objective of this study is to measure aspects of the design process by extracting information from designers' discussion. This research proposes a method to describe the progress of the conceptual design process based on text analysis and information analysis. This study uses word frequencies to represent the evolution of the selection of design alternatives, and information entropy to model the "certainty" of these design choices. An example is given of a real-world team working on the design of a large-scale system architecture problem in an industry setting. Preliminary results of the experiment suggest that this approach to representing the interplay between possible design choices may be useful in determining the progress of early stage design process and understanding the certainty of a design choice.

Keywords: Design Process, Text Analysis, Concept Selection

1 INTRODUCTION
One view of the early stage of design is that it is concerned primarily with the generation and selection of solutions to meet requirements [1]. It has been suggested that 75% of the final production costs are committed in this stage of design [2], giving the preliminary design phase an important role in the overall design process. A common assumption in design is that good design process leads to good design outcome. A range of methods exist for the overall design process, including Quality Function Deployment [3], Axiomatic Design [4], and Systematic Design [5], as well as methods specifically for concept generation such as TRIZ [6]. Many metrics exist to assess the *outcome* of design, such as product or system performance and concept quality [7, 8]. The aim of this study is to examine the viability of metrics for the design *process* by analyzing the language that design teams use. The language that designers use may reflect the design process and analysis of it can provide valuable insights into the design process [9, 10]. The method proposed is a step towards describing the evolution of the conceptual design process over time based on text analysis and information analysis [11, 12]. This study uses word frequencies to represent the evolution of the selection of design alternatives, and information entropy to model the "certainty" of these design choices. This work examines the transcribed discussions of a team designing a large-scale system architecture in an industry setting.

The two research questions explored are:

- *What does the evolution of design alternative choices over time look like?*

  In early stage design, teams typically generate a number of initial concepts and then choose among them, often iterating on the process. What might this process of generation and selection look like graphically speaking? Consider the number of ideas generated during the conceptual design process. It might be expected that the total number of ideas would increase during the concept generation phase and decrease in the concept selection phase as less desirable ideas are weeded out. In the case of the system architecture problem presented in this paper, the focus is on the trade-offs among a smaller number of alternatives rather than on concept generation. It would be anticipated that the volume of concepts would oscillate over time as decisions and trade-offs are made. This study considers the frequency of word occurrences for multiple design alternatives during design discussion to understand the evolution of alternatives in the conceptual design process.

- *How does selection certainty of a design choice change over time?*

Often during conceptual design, there is ambiguity as to whether a design choice is completely settled or “final”. Ambiguity may be valuable in the early stages of design because it can present design freedom. At the same time, understanding the certainty of design choice can help a design progress more efficiently. To gain insight into the certainty of design selections, this study uses information entropy to represent selection certainty. It is expected that design selections become more and more certain as discussions progress over the course of a design process.

2 RELATED WORK

2.1 Text Analysis

There is a great deal of work in computer science that examines ways of extracting and categorizing collections of text documents. Riloff [13] proposed algorithms for automatic text categorization and domain-specific dictionaries. Kitamura et al [14] described a semantic network for representing the functionality of engineering devices. Hearst [15] examined approaches to generate useful groupings. Word similarity and semantic analysis also play a key role in text analysis. Many statistical word similarity measurements have been formulated [16-20], and one well-known application tool is WordNet [21], which is an online lexical reference system used in semantic analysis and information analysis across many domains. While these approaches are important and useful for informing the work described in this study, they are less relevant to representing evolution specifically for design.

2.2 Text Analysis in Design Process

Several researchers have examined design process through surveys [22] and coding of design journals [23]. This study takes a slightly different approach to assessing design process by analyzing the text generated during design. Dong [24, 25] performed Latent Semantic Analysis on design documentation and verbal communication in design teams to express the coherence of design team documents. Correlations were found between more coherent design documentation and better design outcomes. Simpson, et al [26] explored the relationship between the design freedom and the information certainty in the early stage of the design process. Chiu and Shu [27] linked design process language using WordNet as a framework, and built corresponding lexical relationships. Yang, et al [10] examined approaches for the automatic generation of thesauri to improve design information retrieval. Perhaps most closely related to the work described in this paper are the efforts of Song, et al [28], who viewed engineering design as a process of “story telling” and depicted the coherence evolution of the design team in the design process. This study builds on previous design text analysis work, and differs in that it focuses on transcripts of designers rather than with the documentation explicitly created by designers, and also uses both text and information analysis to understand design choice evolution and certainty.

3 METHODS

3.1 Assumptions

This study makes several simplifying assumptions:

Assumption 1: It is assumed that during the design process, what designers say generally corresponds with what they think. This is an implicit assumption of protocol studies of designers [29].

Assumption 2: It is assumed that all major design alternatives for a concept selection problem are largely known *a priori*. While this is probably not true for many early stage design problems, it is a reasonable assumption for a large-scale system re-design problem such as described in this study because they often limit the number of design alternatives under consideration to ones that have been used in the past.

Assumption 3: It is assumed that the more frequently a designer says an alternative, the higher the likelihood it will be chosen for the design, and this relationship between word frequency and selection probability is linear. In other words, people tend to talk more about the design alternative they choose.

3.2 Experimental Study and Data Collection

This study examines a real-world design team working on the design of a large-scale space system architecture. This design team was composed of approximately 15 experienced scientists and engineers of different disciplines working together in a co-located, highly concurrent setting. The team had collaborated on several similar projects in the past. The project took place over three 3-hour sessions spaced out over several weeks. This study focuses on the audio-recorded utterances of one member of the team as he explained his design decision-making process in detail to a novice member of his team. This recording was transcribed into a text document of approximately 28,000 words. All data was time coded. In the transcript, the primary team member talked nearly 85% of the time, and four other members made up the remainder.

3.3 Text and Information Analysis Methods

3.3.1 Design selection evolution

The design project included two design component selection problems, X and Y. For each selection problem, there were 3 alternative candidates. Let X_A, X_B, X_C be the alternative candidates to be chosen for Problem X, and Y_A, Y_B, Y_C be the alternative candidates to be chosen for Y. Much of the design process involved making trade-offs among all the alternative candidates before finally selecting one of three candidates for each design component selection problem (X and Y). The word occurrences of all six alternatives were collected and plotted over time. In this study, the design process is divided into fixed time intervals. These frequencies provided a sense of how much a concept was discussed, and the frequencies of all 3 terms together gave an idea of overall evolution of design choices.

3.3.2 Design selection certainty

For each design alternative in a discussion, the probability that each alternative will be chosen can be calculated based on Assumption 3. Suppose in a selection problem with n alternatives, the design process is divided into m time intervals. Let P_{ij} be the probability with which designers prefer to select the alternative j ($j = 1 \ldots n$) in the i^{th} ($i = 1 \ldots m$) time interval, then P_{ij} can be inferred using maximum likelihood estimation [30] as in Equation 1:

$$P_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} \qquad (1)$$

where f_{ij} is the number of occurrences of the alternative j in the time interval i. With the above probability of occurrences of alternatives, the alternative j is the most likely to occur f_{ij} times if the other alternatives of the same design selection problem occur with certain frequencies.

In a typical design selection process, it might be expected that a choice starts out with relatively low certainty, and becomes more certain towards the end of a project. Information entropy [31] is a measurement of the amount of uncertainty about an event associated with a given probability distribution. The lower the value of information entropy, the more certain the design choice is. Assumption 2 guarantees the completeness of the alternatives in the design selection problem. Let K_i be the selection variable in the i^{th} time interval, whose value can be chosen from the alternative range $\{1, 2, \ldots, n\}$ for a component selection problem with n alternatives. The probability of each alternative of being chosen is known from Equation 1, then the entropy value of the selection variable K_i can be calculated as in Equation 2.

$$entropy(K_i) = -\sum_{j=1}^{n} P_{ij} \log_2 P_{ij} \tag{2}$$

Equation 2 explains how "certain" one design choice is in the i^{th} time interval. Entropy values plotted over time illustrate the changing certainty of the design process. This plot can suggest if designers are sure of the alternative they have chosen already or have still not made a final decision.

4 RESULTS AND DISCUSSION

4.1 Design Selection Evolution

The word occurrences of the six alternatives for the two selection problems in intervals were collected. Figures 1-3 plot the frequencies of the three alternatives of Problem X at intervals of 10 minutes. The dotted vertical lines mark three sessions in which the design process took place. Figure 1 shows how Alternative X_A of Problem X was mentioned often throughout the first period, briefly at the start of the second, and then again throughout the third. This is consistent with a qualitative reading of the transcript which showed that the alternative was "selected" at the end of the first interval, only to be reconsidered in the third. Figure 2 complements Figure 1 and shows that, in the second stage, Alternative X_B was discussed after Alternative X_A had already been chosen. Figure 3 shows that Alternative X_C was discussed constantly throughout the project. From comparison of Figure 1-3, it can be seen that Alternative X_C was discussed more frequently than Alternatives X_A and X_B throughout the design process in Problem X.

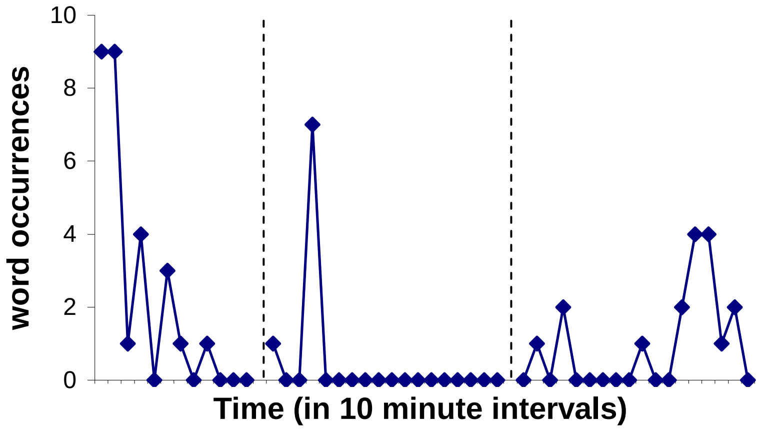

Figure 1. Word Frequencies for Alternative X_A of Problem X

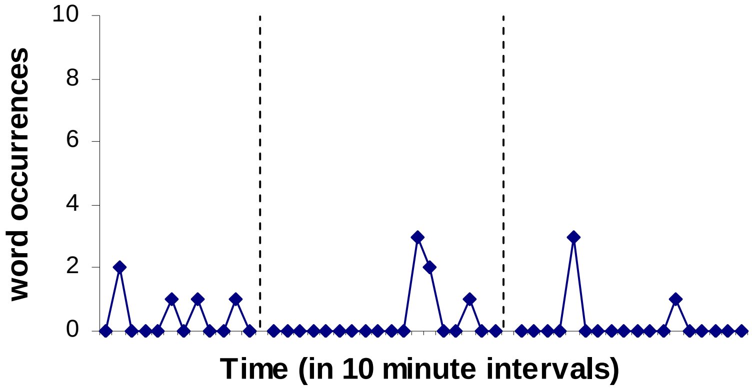

Figure 2. Word Frequencies for Alternative X_B of Problem X

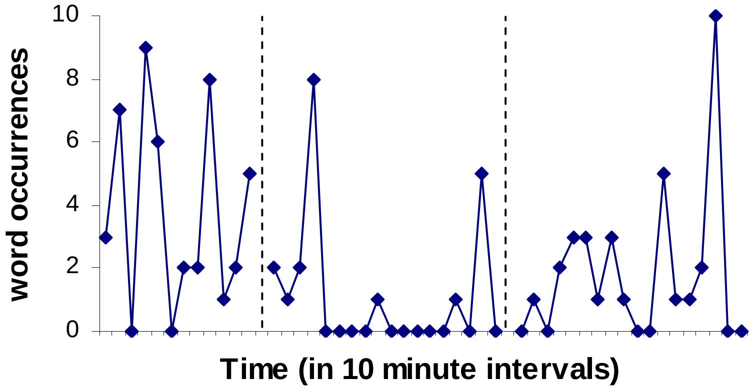

Figure 3. Word Frequencies for Alternative X_C *of Problem X*

Figures 4-6 depict the frequency evolution of the three alternatives of Problem Y. Figure 4 shows that Alternative Y_A was discussed in the first period and then again in the middle of the second period, but it was not mentioned in the third period. Alternatives Y_B and Y_C were only mentioned at the beginning of the first period (Figures 5 and 6). Comparing Figures 4-6 for Problem Y shows that mentions of Alternative Y_A dominated the entire design process. Taken together, these basic frequency plots show graphically when and how alternatives are being considered with respect to other design choices. These plots were surprisingly consistent with qualitative assessments of the transcripts, and seemed to accurately reflect the general decision-making process of the designers. Interestingly, one of the findings of Yang [10] was that design language evolves rapidly over time because of the inherently changing nature of conceptual design process. For example, at the beginning of a project, a designer may refer to a component more generically (a "thing") and then become more specific as design and function become clearer. However, the same pattern of changing terminology was not found here, in part because large scale re-design projects tend be based on well-established "legacy" projects, and the language used to describe the projects likewise tends to be fairly fixed.

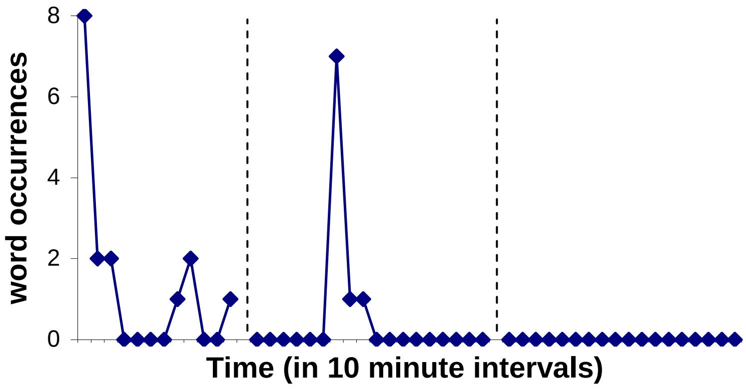

Figure 4. Word Frequencies for Alternative Y_A of Problem Y

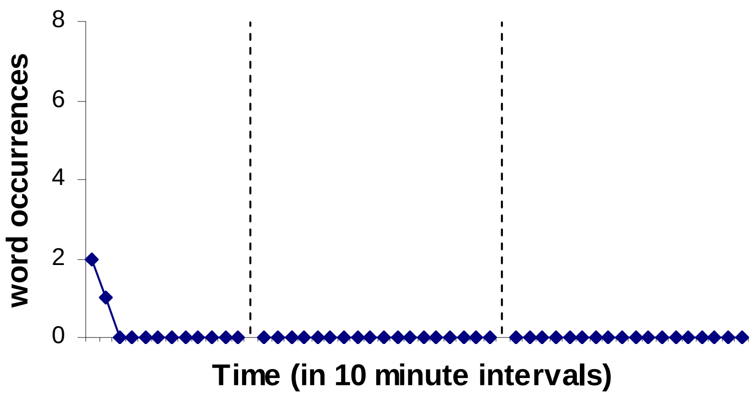

Figure 5. Word Frequencies for Alternative Y_B of Problem Y

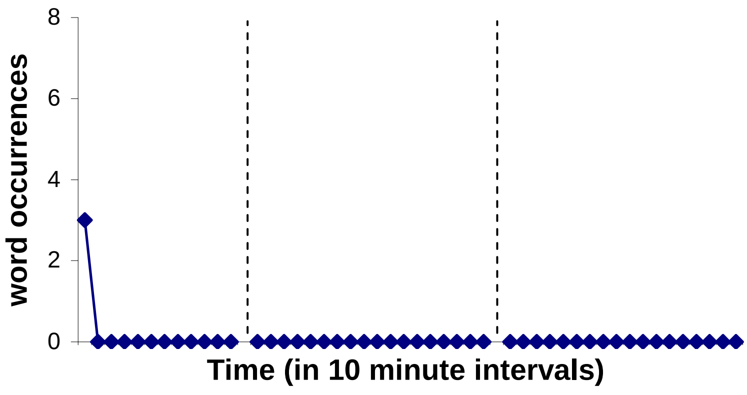

Figure 6. Word Frequencies for Alternative Y_C of Problem Y

4.2 Design selection certainty

The selection probabilities of all six alternatives and information entropies of the two design choices at all time intervals were calculated according to Equation 1 and Equation 2. Figure 7 shows the information entropies of Problem X plotted over time. It suggests that the team actively discussed the alternatives in the first period and perhaps agreed on an alternative in the early stage of the second period where the entropy value is 0, but then diverged again in the third period. This conjecture is partially validated by correspondences with Figures 1-3, and also by a qualitative reading of the original transcript. Figure 8 shows that entropy value of Problem Y is almost 0 except at the beginning of the design process. It means that the designers only discussed this component selection problem early on and were perhaps more certain the alternative should be chosen for this selection problem from the later part of the first period until the end. From the combination of Figure 8 and Figures 4-6, it could be predicted that Alternative Y_A would be chosen for Problem Y, which was validated by examining the original transcript.

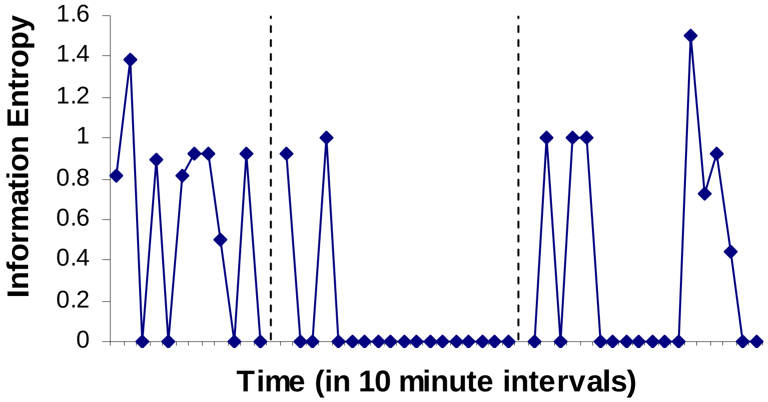

Figure 7. Information Entropy for Problem X

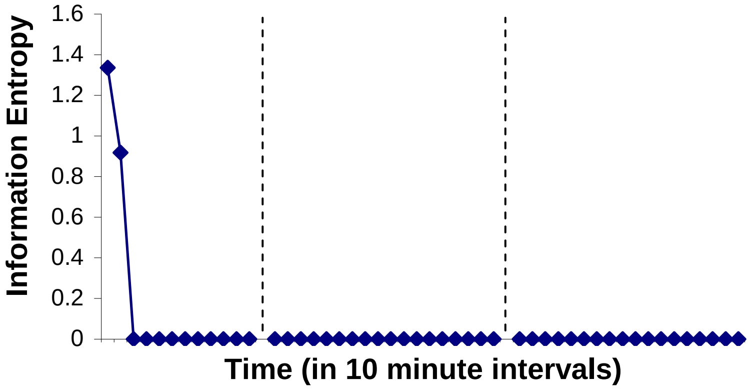

Figure 8. Information Entropy for Problem Y

5 CONCLUSIONS AND FUTURE WORK

The methodology presented in this paper suggests a way to model the design selection process through word frequency analysis and entropy analysis, and may lead to a novel way to understand the nature of design choices over time. In response to the research questions posed in the introduction of this paper, “What does the evolution of design alternative choices over time look like?” and “How does selection certainty of a design choice change over time?," this study gives two case examples that show somewhat different results. It was expected that design alternative choices would oscillate in a large-scale system design problem, and this was true for some design alternatives but not for others. It was anticipated that information entropy would decrease overall with the progress of the design process but the results suggest that patterns are more complex. While this pattern was true for Problem Y, the results for Problem X show that decisions were settled and then became open again for discussion. However, the design process of Problem X did not seem to be well resolved in Figure 7, as the alternative choice was not certain even at the end of the third period. The results of this preliminary study indicate that this approach may be useful in illustrating in detail the progress of early stage design process, but need to be validated and refined on a broader range of design problems.

The method described in this paper makes some simplifying assumptions on how much information about a design is known a priori. Such assumptions are generally true for large system design problem that draw heavily on legacy designs. However, there are several notable engineering design scenarios in which the methods described in this paper may not apply. In particular, the first is when the alternatives are unknown, as in the case of a novel design. This method also assumes discrete design alternatives and it does not apply when the selection range of the design variable is continuous. An example of a continuous variable might be the diameter of a component. There are infinite alternatives in the case when the design variable is continuous, so it is impossible to collect the information of all alternatives.

Future work can take a number of directions. The second assumption can be modified to more closely model real-world problems. Instead of assuming a directly linear relationship, a model can be constructed to describe the more complex relationship between what designers say and what designers think. Applying this model to design documentation or transcripts will provide a more accurate model of the designers' selection process and certainty.

Future research will use concept or semantic weighting instead of frequency to measure the occurrences of the alternatives. The influence of other context-related or semantic-related terms that modify alternatives will be examined. For example, in the context of Alternative A, the word "not" or "no" or "hardly" would be given a negative weight on this alternative. During the design process, designers may use different words to represent the same or the similar meanings. The degree of similarity of different words or phrases should also be considered. Basic lexical relationships among the key words will be determined and the semantic analysis will be employed to extract lexical information such as synonymy, antonymy, hypernymy, hyponymy, meronymy, holonymy, troponymy [32, 33]. Appraisals [34] of the design alternatives will also be quantified to give weightings on the design alternatives. For example, the appraisal word "great" on an alternative in the contextual discussion will give more weighting on that alternative than the appraisal word "good". All these further methods will give more accurate measures of the design alternatives.

In this study, fixed time intervals were specified as a preliminary way to test the approach. The biggest risk of this approach is that there may be too many key alternative words in some intervals, and too few in others to have a meaningful amount of information to examine. This was not a problem for the case study described in this paper as the design team was very intensively concentrated, and the session chosen was not very long. However, for a more complicated and longer design process, fixed time intervals may be insufficient. Future work would involve other approaches to specifying variable time intervals. Transitional words and long time gaps between utterances of alternatives could be taken as indicators as the separators of the time intervals. In the situation with variable time intervals, the word occurrences of alternatives should be normalized over the time duration to make the frequency consistent over the whole design process.

Finally, future work will consider how word frequency and selection certainty measures can be integrated to provide a more comprehensive measure of the design process and describe the design process in a more accurate way.

ACKNOWLEDGEMENTS

The work described in this paper was supported in part by the National Aeronautics and Space Administration under Cooperative Agreement NNA04CL15A and also in part by the National Science Foundation under Award DMI-0547629. The opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors. The authors are indebted to Dr. Leila Meshkat (Caltech/Jet Propulsion Laboratory) for her invaluable insights and comments. The authors gratefully acknowledge helpful discussions with Tomonori Honda (California Institute of Technology), Joshua Scherbinski (Caltech/Jet Propulsion Laboratory), and Dr. Erik Antonsson (California Institute of Technology).

REFERENCES

[1] Pugh, S., *Total Design*. 1991, Wokingham, England: Addison-Wesley.

[2] Ullman, D.G., *The Mechanical Design Process*. Third edition ed. 2003, New York, NY: McGraw-Hill.

[3] Hauser, J. and D. Clausing, *The House of Quality*. Harvard Business Review, 1988. **66**(3): p. 63-73.

[4] Suh, N.P., *The principles of design*. 1990, New York, NY: Oxford University Press.

[5] Pahl, G. and W. Beitz, *Engineering Design - A Systematic Approach*. 1996, Berlin, Heidelberg, New York: Springer-Verlag.

[6] Altshuller, G.S., L. Shulyak, and S. Rodman, *The innovation algorithm : TRIZ, systematic innovation and technical creativity*. 1999, Worcester, Mass: Technical Innovation Center, Inc.

[7] Shah, J.J., S.V. Kulkarni, and N. Vargas-Hernandez, *Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics and Design of Experiments.* Journal of Mechanical Design, 2000. **122**(4): p. 377-384.

[8] Chusilp, P. and Y. Jin, *Impact of Mental Iteration on Concept Generation.* ASME Journal of Mechanical Design, 2006. **128**(1): p. 14-25.

[9] Dong, A., *Concept formation as knowledge accumulation: A computational linguistics study.* AI EDAM (Artificial Intelligence for Engineering Design, Analysis and Manufacturing), 2006. **20**(01): p. 35-53.

[10] Yang, M.C., W. Wood, and M. Cutkosky, *Design information retrieval: a thesauri-based approach for reuse of informal design information.* Engineering with Computers, 2005. **21**(2): p. 177-192.

[11] Suh, N.P., *Axiomatic Design: Advances and Applications*. 2001, Oxford University Press. p. 39-53.

[12] Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**: p. 379-423 & 623-656.

[13] Riloff, E., *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains.* Artificial Intelligence, 1996. **85**(1-2): p. 101-134.

[14] Kitamura, Y., et al. *An Ontology-Based Annotation Framework for Representing the Functionality of Engineering Devices*. in *ASME 2006 International Design Engineering Technical Conferences*. 2006. Philadelphia, PA: ASME.

[15] Hearst, M., *clustering versus faceted categories for informaiton exploration.* Communications of the ACM, 2006. **49**(4): p. 59-61.

[16] Dunning, T., *Accurate methods for the statistics of surprise and coincidence.* Computational Linguistics, 1994. **19**(1): p. 61-74.

[17] Landauer, T.K. and S.T. Dumais, *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.* Psychological Review, 1997. **104**(2): p. 211-240.

[18] Turney, P.D. *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*. in *The 12th European Conference on Machine Learning (ECML 2001)*. 2001. Freiburg, Germany.

[19] Lee, L. *On the effectiveness of the skew divergence for statistical language analysis*. in *Eighth International Workshop on Artificial Intelligence and Statistics*. 2001. Key West, Florida.

[20] Grefenstette, G. *Automatic thesaurus generation from raw text using knowledge-poor techniques*. in *The 9th Annual Conference of the UW Centre for the New OED and text Research*. 1993. Oxford, England.

[21] Miller, G.A., *WordNet: A Lexical Database for English.* Communications of the ACM, 1995.

38(11): p. 39-41.
[22] Brockman, J.B. *Evaluation of Student Design Processes*. in *The 26th Annual Frontiers in Education Conference*. 1996. Salt Lake City, Utah, USA.
[23] Jain, V. and D. Sobek, *Linking design process to customer satisfaction through virtual design of experiments.* Research in Engineering Design, 2006. **17**(2): p. 59-71.
[24] Dong, A., A.W. Hill, and A.M. Agogino, *A Document Analysis Method for Characterizing Team-Based Design Outcomes.* Journal of Mechanical Design, 2004. **126**(3): p. 378-385.
[25] Dong, A., *The latent semantic approach to studying design team communication.* Design Studies, 2005. **26**(5): p. 445-461.
[26] Simpson, T.W., et al. *Metrics for Assessing Design Freedom and Information Certainty in the Early Stages of Design*. in *The 1996 ASME Design Engineering Technical Conferences and Computers in Engineering Conference*. 1996. Irvine, California, USA.
[27] Chiu, I. and L.H. Shu. *Bridging Cross-Domain Terminology for Biomimetic Design*. in *IDETC/CIE 2005 ASME 2005 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. 2005. Long Beach, California USA: ASME.
[28] Song, S., A. Dong, and A.M. Agogino. *Time Variation of Design "Story Telling" in Engineering Design Teams*. in *International Conference on Engineering Design (ICED 03)*. 2003. Stockholm, Sweden.
[29] Cross, N., H. Christiaans, and K. Dorst, eds. *Analysing Design Activity*. First ed. 1996, John Wiley & Sons: West Sussex, England.
[30] Fisher, R.A., *On the mathematical foundations of theoretical statistics.* Philosophical Transactions of the Royal Society, 1922. **222**: p. 309-368.
[31] Shannon, C.E., *A Mathematical Theory of Communication.* Bell Systems Technical Journal, 1948. **27**: p. 379-423, 623-656.
[32] Dong, A. and D. Mclnnes. *Exploring the Relationship between Lexical Behavior and Concept Formation in Design Conversations*. in *ASME 2005 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. 2005. Long Beach, CA, US.
[33] Miller, G.A., et al., *Introduction to WordNet: An On-line Lexical Database.* International Journal of Lexicography, 1990. **3**(4): p. 235-244.
[34] Dong, A., ed. *How Am I Doing? The Language of Appraisal in Design*. Design Computing and Cognition '06 (DCC06), ed. J.S. Gero. 2006, Kluwer: Dordrecht, The Netherlands. 385-404.

Contact: Maria C. Yang
institution/university: Massachusetts Institute of Technology
department: Mechanical Engineering and Engineering Systems Division
street: 77 Massachusetts Avenue
PO Box, City: Cambridge MA 02139-4307
Country: United States
Phone: (617) 324-5592
Fax: (213) 740-1120
e-mail: mcyang@mit.edu
URL: http://www-rcf.usc.edu/~mariayan/