

# Dialogue Generation for Robotic Portraits

Andrew M. Olney  
University of Memphis  
Institute for Intelligent Systems  
aolney@memphis.edu

## Abstract

This paper examines the effectiveness of two domain independent methods of generating dialogue in a conversational robot. The goal is to create interactive dialogues based on the interviews, writings, and biographies of a science fiction author without resorting to domain specific knowledge engineering. Judges read conversations between the author and an interviewer, where the last utterance is either a continuation of the original dialogue, randomly generated, or generated using a combination of shallow semantic, statistical, and syntactic techniques. By rating the last utterance on several scales related to Grice's conversational maxims and the Bystander Turing Test, the expert judges enable comparison amongst the dialogue generation methods for this specific domain. The dialogue generation methods were then incorporated into a robotic portrait of the science fiction author.

## 1 Introduction

How can we best hold a natural conversation with a machine? The difficulties associated with an unrestricted natural language interface are so severe that the inventor of the computer, Alan Turing, proposed it be the condition upon which computers were deemed to have "intelligence" [Turing, 1950]. Turing's original proposal is called an imitation game, where a human seated at a computer terminal has two conversations via the keyboard and screen. One of the conversational partners is a human, the other a machine. If the human seated at the terminal is unable to distinguish between them, the machine has passed the Turing Test. Notwithstanding the validity of determining intelligence based on behavior, this technique, or the Bystander variant, has been used in evaluating natural language generation [Jurafsky and Martin, 2000; Person *et al.*, 2002]. It has also fired the public's imagination in the annual Loebner Prize competition established to encourage the development of machines that could pass such a test.

However, many previous attempts at the Turing Test have dubious merit. Shieber [1994] in a critique of the Loebner competition concludes that the scope of the endeavor is too much for the current state of artificial intelligence. The

weight of his argument stands on the methods used by the winning programs: the same kind of trickery first applied almost 40 years ago by ELIZA [Weizenbaum, 1966]. Programs such as ELIZA, PARRY [Colby, 1981], and the winner of the first Loebner competition [Weintraub, 1991] all select a conversational mode that masks deficiencies in the natural language interface. ELIZA imitates a Rogerian psychotherapist, which by definition is supposed to 'reflect' the patient/user's statements:

PATIENT: I dislike you.

ELIZA: Does it please you to dislike me?

PARRY likewise reflects the user's utterances, this time using the idiom of a paranoid schizophrenic:

USER: I like to play football.

PARRY: Why do you keep talking about football?  
You are really getting on my nerves.

Likewise, Weintraub's program used the idiom of "whimsical conversation". All of these programs use simple string matching to extract a phrase from the user's utterance and include that string in the program's utterance. The idioms used by the programs limit the user's expectations beyond a turn, and the idioms thus mask the simplicity of the algorithm. It's been reported that psychiatrists have believed PARRY's transcripts belong to a real paranoid schizophrenic, but the scientific merit of such an achievement is highly questionable [Shieber, 1994].

Although Shieber [1994]'s analysis reveals the enormous gap between winning the Loebner prize and possessing "real" intelligence, it is nonetheless significant that modern reimplementations of the pattern matching strategy using AIML [Wallace, 2005] are able to hold a realistic conversation on a number of topics. The ability of such a "chatterbot" to hold a realistic conversation can be useful in some contexts, e.g. video game entertainment, product marketing, and education.

Unfortunately, the effort involved in scaling up the pattern matching approach is enormous: freely available AIML implementations have upwards of 43,000 unique patterns, which together with the corresponding response of the chatterbot must be crafted by hand. Existing chatterbot packages [Foundation, 2005] mitigate this problem by allowing online users to interact with the chatterbot in a kind of reinforce-

ment learning, which is somewhat less effortful than scripting AIML by hand. Overall however, the AIML scripting approach suffers from the same weaknesses as any other knowledge engineering methodology: all of the knowledge must be engineered by hand, and the knowledge is usually domain dependent.

In contrast, this paper investigates domain independent generation of dialogue using techniques from shallow semantics, statistics, and syntax, all of which are either domain independent or use unsupervised learning. Even though the techniques are domain independent, some domain must be chosen for the purposes of evaluation. Therefore we have selected Philip K. Dick (PKD), a science fiction writer of the 20th century, as the dialogue persona to emulate. That is, the chatterbot must be able to respond both with the relevant information and with the relevant PKD style, just as if PKD himself had responded. The task is to utilize a collection of interviews, biographies, and novels written by PKD as a source for generating dialogue that has relevant response information and a stylistically appropriate answer. The best such dialogue generation techniques have been loaded into the PKD Android, a robotic portrait of Philip K. Dick.

## 2 Biographical Dialogue Generation

Depending on the unit size of generation, e.g. a paragraph, a sentence, a phrase, different problems arise. At the paragraph level, the generation task is essentially information retrieval (IR): given a context of dialogue (a query) find the best matching paragraph in the collection of interviews. This approach has the advantage that the paragraph returned is guaranteed to be a well formed PKD response and therefore to have his style. The disadvantage with this approach is that it is unlikely that an existing paragraph is totally appropriate, unless the context of the conversation closely matches that of a previous interview.

At the sentence level, the generation task is essentially IR with multiple document summarization: given a context of dialogue find the best matching documents, find the most appropriate summary sentences, and sequence them. The advantage of this approach is that the content and style of a response can be more closely tailored to the current context of the conversation. The disadvantage is that now we must consider the ordering of the sentences and the relationship between the sentences. For example, the cohesion between sentences may no longer match the PKD style, and anaphora between sentences need to be resolved.

Finally, at the phrase level, the generation task combines information retrieval, summarization, and parsing: retrieve the most relevant documents, find summary sentences, extract the phrases from these and recombine them into a syntactically and semantically well-formed response. Clearly the advantages are even more flexibility in the response, so that the chatterbot's reply may very closely match the context of the conversation. The disadvantage is that on top of the disadvantages inherited from the other two levels, the phrase approach must solve the problem of phrase order, e.g. matching verbs with appropriate semantic and syntactic arguments. Given the existing state of the natural language generation field, this is

likely the point of diminishing returns at which the flexibility of the unit size is offset by the complexity of combining like-sized units together.

Our approach is distinguished from previous statistical approaches to natural language generation in that we use no underlying meaning representation. For example, Langkilde & Knight [1998a] and Langkilde & Knight [1998b], who pioneered the first statistical approach to NLG using bigrams, use an underlying abstract meaning representation (AMR) a kind of propositional structure, as the semantic specification of the desired generated text. This kind of semantic specification is one way to determine rough word order in a generated sentence, e.g. actor/actee roles. Our approach on the other hand, gets all of its semantics from latent semantic analysis (LSA) a statistical technique for modeling world knowledge. Not only can LSA be used for information retrieval [Dumais, 1993] but it also can be used to measure the cohesion between sentences and paragraphs [Foltz *et al.*, 1998] a useful property for sequencing sentences.

### 2.1 Paragraph: Information Retrieval

Latent semantic analysis [Landauer and Dumais, 1997; Dumais, 1993; Manning and Schutze, 1999] is a statistical technique that represents the similarity between collections of words as a cosine between vectors, and LSA can be used for information retrieval. The process begins with the collection of text into a single corpus. A matrix is created from the corpus, having one row for each unique word in the corpus and one column for each document or paragraph. The cells of the matrix consist of a simple count of the number of times word  $i$  appeared in document  $j$ . Since many words do not appear in any given document, the matrix is often sparse. Weightings are applied to the cells that take into account the frequency of word  $i$  in document  $j$  and the frequency of word  $i$  across all documents, such that distinctive words that appear infrequently are given the most weight.

The key to the process is singular value decomposition (SVD), a technique that creates an approximation of the original word by document matrix. After SVD, the original matrix is equal to the product of three matrices, word by singular value, singular value by singular value, and singular value by document. The size of each singular value corresponds to the amount of variance captured by a particular dimension of the matrix. Because the singular values are ordered in decreasing size, it is possible to remove the smaller dimensions and still account for most of the variance. The approximation to the original matrix is optimal, in the least squares sense, for any number of dimensions one would choose. In addition, the removal of smaller dimensions introduces linear dependencies between words that are distinct only in dimensions that account for the least variance. Consequently, two words that were distant in the original space can be near in the compressed space, supporting the inductive machine learning and knowledge acquisition reported in the literature [Landauer and Dumais, 1997]. The final step is to compare two collections of words by creating two vectors. Each word is associated with a row vector in the matrix, and the vector of a collection is simply the sum of all the row vectors of words in that collection. Vectors are compared geometrically using cosine.

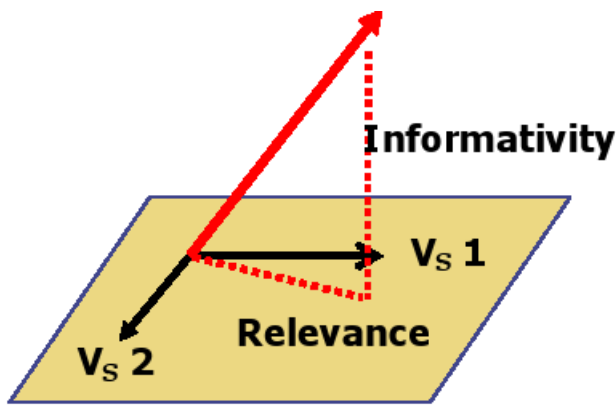


Figure 1: An orthonormal basis

For the PKD task, the interviewer’s utterance is used as a query into a database containing conversations between interviewers and PKD. The  $N$  highest matching documents are checked to ensure that they are PKD responses. If one of the highest matching documents is an interviewer utterance, then the PKD response to that interviewer utterance is used. This method ensures that the PKD response is topically cohesive with the interviewer’s utterance.

## 2.2 Sentence: Summarization

Cohesion can be measured by comparing the LSA cosines of two successive sentences or paragraphs [Foltz *et al.*, 1998]. However, cohesion is a crude metric: repetitions of a single sentence will be highly cohesive (cosine of 1) but conversationally inappropriate since no new information is introduced. A variation of the LSA algorithm using orthonormalized vectors provides two new measures, informativity and relevance, which can detect how much new information is added and how relevant it is in a context [Hu *et al.*, 2003]. The essential idea is to represent context by an orthonormalized span of vectors, one vector for each utterance. The span is a subspace of the higher dimensional LSA space, in the same way as a plane or line is a subspace of 3D space. The span is created by projecting each utterance onto the span of previous utterances using a method known as the Gram-Schmidt process [Anton, 2000]. Since each vector in the span is orthogonal, the span represents all linear combinations of what has been previously said. For example, in Figure 1, a new utterance creates a new vector that can be projected to the span, forming a triangle. The base of the triangle lies within the plane created by  $V_s 1$  and  $V_s 2$ , the two previous vectors projected into the span. The leg of the triangle that lies within the span indicates the relevance of the recent utterance to the span; the perpendicular leg indicates new information. Accordingly, a repeated utterance would have complete relevance but zero new information. This discrimination is key in any attempt to use LSA to generate natural language.

The orthonormal basis can theoretically be used to create a summary of a group of paragraphs, although this technique has never been used before. Using LSA to select  $N$  documents, suppose that these  $N$  documents are segmented into

$M$  sentences. These  $M$  sentences can be used to form  $M$  orthonormal bases, each consisting of  $M - 1$  sentences. When the excluded sentence is projected into each of these orthonormal bases, two measures are returned, informativity and relevance. The excluded sentence with the lowest informativity, i.e. whose meaning was mostly captured by all the other sentences, is discarded, and the process begun again with  $M - 1$  sentences. The stopping criteria for such an algorithm can either be the number of sentences in the summary or a threshold for informativity.

To preserve the conversational stylistics of PKD, a set of  $N$  turns was summarized to  $K$  sentences. The sentence that appeared closest to the beginning of its respective turn was selected as the starter sentence. Stylistically, this technique addresses the issue of discourse markers and conversational openings. The number of sentences in the turn associated with the starter sentence (starter turn) was used as the number of target sentences to have in the generated turn, and the cohesion relationships between sentences in that turn was used as the target for the cohesion of sentences in the generated turn. For example, first the starter sentence is projected into the span. Second, each of the remaining  $K - 1$  sentences is projected into the span, and the sentence whose relevance/informativity most closely mirrors the relevance/informativity of the second sentence in the starter turn is selected. The second step repeats for each additional sentence to be generated. This technique ensured that the cohesion relationships between sentences in the generated turn matched as closely as possible with the cohesion relationships in the starter turn (and therefore with a natural PKD turn).

## 3 Method

### 3.1 Independent Variables

Two independent variables were manipulated. The first independent variable is the method of generating natural language, either continuing actual human dialogue, randomly, or using NLG techniques. The conditions of using actual dialogue and random dialogue provide control conditions and hypothetically represent the upper and lower bounds to which to compare the generation algorithm. The second independent variable is unit of generation (paragraph or sentence). Only the last utterance of the dialogue [Person *et al.*, 2002] is generated.

### 3.2 Dependent variables

Five dependent variables were measured. For each conversation judges were asked to respond to five questions. The first question is the Bystander Turing Test question, i.e. “Was the last utterance generated by a computer?” The other four questions directly correspond with the Gricean Conversational Maxims (Grice, 1975) of Quantity, Quality, Relevance, and Manner:

- The “PKD” answer is an appropriate length
- The “PKD” answer is accurate (i.e. contains no falsehoods)
- The “PKD” answer is relevant to the non PKD sentence preceding it

- The “PKD” answer is clear, unambiguous, and orderly

All ratings are on a 6-point scale:

- Definitely not
- Probably not
- Not sure but guess not
- Not sure but guess yes
- Probably yes
- Definitely yes

### 3.3 Judges

Two judges participated. Judge A had low knowledge of PKD. Judge B had an above average knowledge of PKD.

### 3.4 Materials

The methods for generating natural and random conversations are defined as follows. Approximately 8500 sentences from PKD interviews were extracted from their text sources and used to build a conversation database. Each natural conversation was created by randomly selecting an interviewer question in the database and then selecting PKD’s response to that question. Random conversations were generated by randomly selecting a question by an interviewer and then randomly selecting a PKD response. In the case of randomly generated sentences, the first sentence was selected from the random PKD response. If the response originally had more than one sentence, then that number of additional sentences was selected from the database, taking care not to select the same sentence twice. Therefore the random sentence condition created PKD answers where the sentences were from various turns, and the number of sentences was consistent with the first turn selected.

The algorithm-generated conversations were created by creating an LSA space of 23 scanned books and the text of the conversation database. New software was written to implement the techniques mentioned above for LSA based information retrieval and orthonormal basis summarization for the respective text units of paragraph and sentence. Note that for LSA based IR, the second highest document was chosen, as opposed to the first (which would otherwise give the actual PKD response).

### 3.5 Procedure

Judge evaluation consisted of rating one packet containing 60 interviewer/PKD pairs. The method variable has three levels, natural, random, and generated, the unit variable has two levels, paragraph and sentence, and each condition had ten exemplars each,  $3 \times 2 \times 10 = 60$ . The pairs were randomly ordered. The judges were asked to rate the conversation on 5 six-point scales, indicating the Bystander Turing Test and the four Gricean Maxims. The following is an example dialogue turn used in the experiment:

INTERVIEWER: did he take you for real oh my god he really thought you rode horses huh  
 DICK: oh well christ i have no idea what he thought

Measure	Natural	Random	Generated	Total
<b>The “PKD” answer was generated by a computer</b>				
Paragraph	2.45	3.05	3.40	8.90
Sentence	3.00	4.35	3.65	11.00
Total	5.45	7.40	7.05	
<b>The “PKD” answer is an appropriate length</b>				
Paragraph	4.30	3.55	4.00	11.85
Sentence	4.30	4.00	3.85	12.15
Total	8.60	7.55	7.85	
<b>The “PKD” answer is accurate</b>				
Paragraph	4.65	3.50	4.25	12.40
Sentence	4.05	3.45	3.90	11.40
Total	8.70	6.95	8.15	
<b>The “PKD” answer is relevant</b>				
Paragraph	4.65	2.50	3.80	10.95
Sentence	3.95	2.50	3.15	9.60
Total	8.60	5.00	6.95	
<b>The “PKD” answer is clear</b>				
Paragraph	4.95	3.85	4.40	13.20
Sentence	3.95	2.95	3.45	10.35
Total	8.90	6.00	7.85	

Figure 2: Means for dependent measures

## 4 Results

Five  $3 \times 2$  repeated measures ANOVAs were performed to determine the effect of Generation Method (Natural vs. Random vs. NLG) and Unit Size (Paragraph vs. Sentence) on the subject’s ratings of the five questions. The means for the four measures are reported in Figure 2.

### 4.1 Question 1: The “PKD” answer was generated by a computer

The  $3 \times 2$  ANOVA for the “PKD answer was generated by a computer” question indicated no interaction between the method of generation and the unit size. However, the method of generation was significant ( $F = 5.78, p = .01$ ) and the unit size was also significant ( $F=8.94, p = .02$ ). Pairwise comparison of the methods revealed that the natural method was significantly different from the random condition ( $p = .002$ ) and from the generated condition ( $p = .04$ ), but there is no statistical difference on this question between the random and generated conditions. Analysis of the associated means shows that the natural responses were perceived as more humanlike than the other two. Pairwise comparison of the units revealed that the paragraph unit is significantly different from the sentence unit ( $p = .02$ ). Analysis of the associated means shows that the paragraphs were perceived as more humanlike than sentences. The difference between the means for random paragraphs and random sentences almost leads to a significant interaction ( $F = 2.82, p = .09$ ). Interestingly, the difference between the means for natural paragraph and natural sentence is greater than the difference between generated paragraph

and generated sentence, indicating that the orthonormal basis method of generating a response is doing well compared to the IR approach on this measure.

#### 4.2 Question 2: The “PKD” answer is an appropriate length

The 3 x 2 ANOVA for the “PKD answer is an appropriate length” question indicated no significant interaction between the method of generation and the unit size. There was no significant difference between methods of generation and no significant difference between unit sizes.

#### 4.3 Question 3: The “PKD” answer is accurate

The 3 x 2 ANOVA for the “PKD answer is accurate” question indicated no interaction between the method of generation and the unit size. However, the method of generation was significant ( $F = 3.64$ ,  $p = .05$ ) and the unit size was also significant ( $F = 6.43$ ,  $p = .03$ ). Pairwise comparison of the methods revealed that the natural method was significantly different from the random condition ( $p = .03$ ), but there is no statistical difference on this question between the generated condition and the other two conditions. Analysis of the associated means shows that the natural responses were perceived as more accurate than random responses, and the nonsignificant trend is that natural responses are more accurate than generated which are more accurate than random. Pairwise comparison of the units revealed that the paragraph unit is significantly different from the sentence unit ( $p = .03$ ). Analysis of the associated means shows that the paragraphs were perceived as more accurate than sentences. The difference between the means for random paragraphs and random sentences almost leads to a significant interaction ( $F = 2.82$ ,  $p = .09$ ). Once again, the difference between the means for natural paragraph and natural sentence is greater than the difference between generated paragraph and generated sentence, indicating that the orthonormal basis method of generating a response is doing well compared to the IR approach on this measure.

#### 4.4 Question 4: The “PKD” answer is relevant

The 3 x 2 ANOVA for the “PKD answer is relevant” question indicated no interaction between the method of generation and the unit size. However, the method of generation was significant ( $F = 14.07$ ,  $p = .0001$ ) and the unit size was also significant ( $F = 5.46$ ,  $p = .04$ ). Pairwise comparison of the methods revealed that the random method was significantly different from the natural condition ( $p = .0001$ ) and from the generated condition ( $p = .02$ ), but there is no statistical difference on this question between the natural and generated conditions. Analysis of the associated means shows that the natural and generated responses were perceived as more relevant than random responses. Pairwise comparison of the units revealed that the paragraph unit is significantly different from the sentence unit ( $p = .04$ ). Analysis of the associated means shows that the paragraphs were perceived as more relevant than sentences. Again, the difference between the means for natural paragraph and natural sentence is smaller than the difference between generated paragraph and generated sentence, indicating that the orthonormal basis method of generating a

response is doing well compared to the IR approach on this measure.

#### 4.5 Question 5: The “PKD” answer is clear

The 3 x 2 ANOVA for the “PKD answer is clear” question indicated no interaction between the method of generation and the unit size. The method of generation was not significant ( $F = 3.02$ ,  $p = .07$ ) and the unit size was significant ( $F = 21.49$ ,  $p = .001$ ). Analysis of the means of method shows the nonsignificant trend of natural as more clear than generated, which is more clear than random. Pairwise comparison of the units revealed that the paragraph unit is significantly different from the sentence unit ( $p = .001$ ). Analysis of the associated means shows that the paragraphs were perceived as more accurate than sentences.

## 5 Discussion

With the exception of the first question, “The PKD answer was generated by a computer,” all answers to the questions show a trend in which the generated answer is between random and natural. This trend meets our expectations that the generation techniques will provide a useful approximation of the natural answer. Moreover, on all questions except the first, there is no statistically significant difference between the natural method and the generated method. This is an extremely positive result. It means that for all of Grice’s conversational maxims, there is no perceptible difference between the natural and generated methods. Therefore these results show that the generated method performs as an appropriate conversational partner for one turn. If it can be determined what is cuing the judges on the first question to rate the generated responses as more computerlike, the system should be able to generate responses that are not significantly different from the natural response on any of the five measures.

With the exception of the second question, “The PKD response was of an appropriate length” which was not significant, the effect of unit size was significant on all questions. The direction of the significant difference indicates that paragraphs are considered more humanlike and more conversationally appropriate than sentences. This result suggests that a backoff strategy is appropriate for this task. In such a backoff strategy, the paragraph method would be used preferentially before the sentence method. The exact decision criteria for this strategy has been determined in post hoc tests, in which a regression model was fit to metrics such as cohesion, informativity, and relevance. It is also interesting to note that the difference between the means for sentence and paragraph for the generated method on questions 1, 3, and 4, was less than the same difference for the natural method. This indicates that the summarization method using an orthonormal basis is doing fairly well, and suggests that some of the significant differences between sentence and paragraph are a result of the greatly differing means in the random sentence and random paragraph conditions.

Overall, these results show that shallow, domain-independent, and unsupervised learning techniques can be used to create a conversationally appropriate dialogue between a computer and a person, i.e. one that follows the



Figure 3: A robotic portrait of PKD

Gricean Maxims. This is particularly interesting because it shows that knowledge engineering techniques so common in natural language generation research are not necessary to make a conversationally appropriate dialogue. The issue of passing the Bystander Turing Test, however, still needs to be addressed.

These results were incorporated into a robotic portrait of PKD [Hanson *et al.*, 2005] shown in Figure 3. Although a full description of the PKD Android is beyond the scope of this paper, the following gives a brief overview of its implemented dialogue generation system. Because of processing constraints, the final dialogue generation system used the paragraph generation method described in this paper together with a chatterbot for stock questions, e.g. “What is your name”. The paragraph method generated 20 responses and the chatterbot generated 1 response to each user input. Responses were then ranked as to their probability of being a good response using the metrics of cohesion, relevance, and informativity described above. If the best response had a low probability of being an end of turn response, based on the probability of its last word being the end of a turn, the next best response was appended onto the end, iteratively. This final answer was then spoken by the PKD Android, with corresponding lip synchronozation and facial expression. An example conversation follows:

PKD Android: Well the I-Ching gives advice beyond the particular advice that transcends the immediate situation. The answers have a universal quality. For instance, the mighty are humbled, and the humbled are raised. If you use the I-Ching long enough and continually enough it will begin to change and shape you as a person. It will make you into a Taoist, whether or not you have ever heard

the word, whether or not you want to be. Well...

Human: Don't you think Taoism is too easy a religion for a robot? To illogical?

PKD Android: I think the powers actually exist. I think they are real. Too: are you talking about an animal, vegetable, or mineral?

Human: Taoism doesn't have to do so much with animals, vegetables, or minerals, but ideas that you can't prove. How can a robot think about something that it can't prove?

PKD Android: What makes it impossible?

Human: What makes it impossible to prove Taoism? It's too abstract.

PKD Android: Ask the Open Directory about it. It being what.

Human: It being Taoism...

## 6 Conclusion

This paper examined the effectiveness of several domain independent methods of generating dialogue, and introduced the use of the orthonormal basis for summarization. By mining a large collection of text written by and about Philip K Dick, LSA and the orthonormal basis for summarization were used to generate PKD-style answers to randomly selected interviewer questions. Judges read conversations between PKD and an interviewer, where the last utterance was either a continuation of the original dialogue, randomly generated, or generated using LSA or the orthonormal basis for summarization. Analysis of the judges ratings showed that although the techniques introduced do not allow a computer to pass the Bystander Turing Test, they do create a conversational partner that is indistinguishable from PKD on all of Grice's conversational maxims. This result indicates that natural language generation methods do not have to include knowledge engineering and that domain independent and unsupervised techniques can be used to create an effective conversational partner.

## References

- [Anton, 2000] H. Anton. *Elementary linear algebra*. John Wiley, New York, 2000.
- [Colby, 1981] K. M. Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4(4):515–560, 1981.
- [Dumais, 1993] S. Dumais. LSI meets TREC: A status report. In *Proceedings of the First Text Retrieval Conference*, pages 137–152. NIST, 1993.

- [Foltz *et al.*, 1998] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307, 1998.
- [Foundation, 2005] ALICE Artificial Intelligence Foundation, 2005.
- [Hanson *et al.*, 2005] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou. Upending the uncanny valley. In *Proceedings of the Twentieth National Conference on Artificial Intelligence and the Seventeenth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 1728–1729, Menlo Park, CA, 2005. AAAI Press.
- [Hu *et al.*, 2003] X. Hu, Z. Cai, M. Louwerse, A. Olney, P. Penumatsa, and A. Graesser. An improved LSA algorithm to evaluate contributions in student dialogue. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 2003.
- [Jurafsky and Martin, 2000] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, 2000.
- [Landauer and Dumais, 1997] T. Landauer and S. Dumais. A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
- [Langkilde and Knight, 1998a] Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL*, pages 704–710, 1998.
- [Langkilde and Knight, 1998b] Irene Langkilde and Kevin Knight. The practical value of N-grams in derivation. In Eduard Hovy, editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 248–255, New Brunswick, New Jersey, 1998. Association for Computational Linguistics.
- [Manning and Schutze, 1999] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, 1999.
- [Person *et al.*, 2002] N. K. Person, M. Erkel, and A. C. Graesser. Autotutor passes the bystander Turing test. In *Proceedings of E-Learn*, 2002.
- [Shieber, 1994] S. M. Shieber. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery*, 37(6):70–78, 1994.
- [Turing, 1950] A. M. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.
- [Wallace, 2005] R. Wallace. <http://www.alicebot.org/TR/2005/WD-aiml/>, 2005.
- [Weintraub, 1991] J. Weintraub. <http://www.loebner.net/Prizef/weintraub-bio.html>, 1991.
- [Weizenbaum, 1966] J. Weizenbaum. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9:36–45, 1966.