Utterance Classification in AutoTutor

Andrew Olney **Max Louwerse** **Eric Matthews** **Johanna Marineau** **Heather Hite-Mitchell** **Arthur Graesser**

Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

Abstract

This paper describes classification of typed student utterances within AutoTutor, an intelligent tutoring system. Utterances are classified to one of 18 categories, including 16 question categories. The classifier presented uses part of speech tagging, cascaded finite state transducers, and simple disambiguation rules. Shallow NLP is well suited to the task: session log file analysis reveals significant classification of eleven question categories, frozen expressions, and assertions.

1 Introduction

AutoTutor is a domain-portable intelligent tutoring system (ITS) with current versions in the domains of physics and computer literacy (Graesser et al. 1999; Olney et al. 2002). AutoTutor, like many other ITSs, is an intersection of applications, including tutoring, mixed-initiative dialogue, and question answering. In each of these, utterance classification, particularly question classification, plays a critical role.

In tutoring, utterance classification can be used to track the student's level of understanding. Contribution and question classifications can both play a role: contributions may be compared to an expected answer (Graesser et al. 2001) and questions may be scored by how "deep" they are. For example, The PREG model (Otero and Graesser 2001) predicts under what circumstances students will ask "deep" questions, i.e. those that reveal a greater level of cognitive processing than who, what, when, or where questions. A student who is only asking shallow questions, or no questions at all, is predicted by PREG to not have a situation-level understanding (van Dijk and Kintsch 1983) and thus to learn less and forget faster. The key point is that different metrics for tracking student understanding are applicable to questions and contributions. Distinguishing them via classification is a first step to applying a metric.

In mixed-initiative dialog systems, utterance classification can be used to detect shifts in initiative. For example, a mixed-initiative system that asks, "Where would you like to travel", could respond to the question, "Where can I travel for $200?" (Allen 1999) by giving a list of cities. In this example, the user is taking the initiative by requesting more information. In order to respond properly, the system must detect that the user has taken initiative before it can respond appropriately; otherwise it might try to interpret the user's utterance as a travel destination. In this sense, questions mark redirection of the dialogue, whereas contributions are continuations of the dialogue. In order for a user to redirect the dialogue and thus exercise initiative, a mixed-initiative system must be able to distinguish questions and contributions.

Question classification as early as Lehnert (1978) has been used as a basis for answering questions, a trend that continues today (Voorhees 2001). A common feature of these question-answering systems is that they first determine the expected answer type implicit in the question. For example, "How much does a pretzel cost" might be classified according to the answer type of MONEY or QUANTITY. Knowledge of the expected answer type can be used to narrow the search space for the answer, either online (Brill et al. 2001) or in a database (Harabagiu et al. 2000). Accordingly, question answering calls for a finer discrimination of question types as opposed to only distinguishing questions from contributions.

AutoTutor uses utterance classification to track student progress, to determine initiative, and to answer questions. By virtue of being embedded in AutoTutor, the utterance classifier presented here has an unusual set of constraints, both practical and theoretical. On the practical side, AutoTutor is a web-based application that performs in real time; thus utterance classification must also proceed in real time. For that reason, the classifier uses a minimum of resources, including part of speech tagging (Brill 1995; Sekine and Grishman 1995) and cascaded finite state transducers defining the categories. Theoretically speaking, AutoTutor must also recognize questions in a meaningful way to both question answering and tutoring. The question taxonomy utilized, that of Graesser et al (1992), is an extension of Lehnert's (1978) taxonomy for question answering and has been applied to human tutoring (Graesser et al. 1992; Graesser and Person 1994).

This paper outlines the utterance classifier and quantifies its performance. In particular, Section 2 presents AutoTutor. Section 3 presents the utterance taxonomy. Section 4 describes the classifier algorithm. Section 5 delineates the training process and results. Section 6 presents evaluation of the classifier on real AutoTutor sessions. Section 7 concludes the paper.

2 AutoTutor

AutoTutor is an ITS applicable to any content domain. Two distinct domain applications of AutoTutor are available on the Internet, for computer literacy and conceptual physics. The computer literacy AutoTutor, which has now been used in experimental evaluations by over 200 students, tutors students on core computer literacy topics covered in an introductory course, such as operating systems, the Internet, and hardware. The topics covered by the physics AutoTutor are grounded in basic Newtonian mechanics and are of a similar introductory nature. It has been well documented that AutoTutor promotes learning gains in both versions (Person et al. 2001).

AutoTutor simulates the dialog patterns and pedagogical strategies of human tutors in a conversational interface that supports mixed-initiative dialog. AutoTutor's architecture is comprised of seven highly modular components: (1) an animated agent, (2) a curriculum script, (3) a speech act classifier, (4) latent semantic analysis (LSA), (5) a dialog move generator, (6) a Dialog Advancer Network, and (7) a question-answering tool (Graesser et al. 1998; Graesser et al. 2001; Graesser et al. 2001; Person et al. 2000; Person et al. 2001; Wiemer-Hastings et al. 1998).

A tutoring session begins with a brief introduction from AutoTutor's three-dimensional animated agent. AutoTutor then asks the student a question from one of topics in the curriculum script. The curriculum script contains lesson-specific tutor-initiated dialog, including important concepts, questions, cases, and problems (Graesser and Person 1994; Graesser et al. 1995; McArthur et al. 1990; Putnam 1987). The student submits a response to the question by typing and pressing the "Submit" button. The student's contribution is then segmented, parsed (Sekine and Grishman 1995) and sent through a rule-based utterance classifier. The classification process makes use of only the contribution text and part-of-speech tag provided by the parser.

Mixed-initiative dialog starts with utterance classification and ends with dialog move generation, which can include question answering, repeating the question for the student, or just encouraging the student. Concurrently, the LSA module evaluates the quality of the student contributions, and in the tutor-initiative mode, the dialog move generator selects one or a combination of specific dialog moves that is both conversationally and pedagogically appropriate (Person et al 2000; Person et al. 2001). The Dialog Advancer Network (DAN) is the intermediary of dialog move generation in all instances, using information from the speech act classifier and LSA to select the next dialog move type and appropriate discourse markers. The dialog move generator selects the actual move. There are twelve types of dialog move: Pump, Hint, Splice, Prompt, Prompt Response, Elaboration, Summary, and five forms of immediate short-feedback (Graesser and Person 1994; Graesser et al. 1995; Person and Graesser 1999).

3 An utterance taxonomy

The framework for utterance classification in Table 1 is familiar to taxonomies in the cognitive sciences (Graesser et al. 1992; Graesser and Person 1994). The most notable system within this framework is QUALM (Lehnert 1978), which utilizes twelve of the question categories. The taxonomy can be divided into 3 distinct groups, questions, frozen expressions, and contributions. Each of these will be discussed in turn.

The conceptual basis of the question categories arises from the observation that the same question may be asked in different ways, e.g. "What happened?" and "How did this happen?" Correspondingly, a single lexical stem for a question, like "What" can be polysemous, e.g. both in a definition category, "What is the definition of gravity?" and metacommunicative, "What did you say?" Furthermore, implicit questions can arise in tutoring via directives and some assertions, e.g. "Tell me about gravity" and "I don't know what gravity is." In AutoTutor these information seeking utterances are classified to one of the 16 question categories.

The emphases on queried concepts rather than orthographic forms make the categories listed in Table 1 bear a strong resemblance to speech acts. Indeed, Graesser et al. (1992) propose that the categories be distinguished in precisely the same way as speech acts, using semantic, conceptual, and pragmatic criteria as opposed to syntactic and lexical criteria. Speech acts presumably transcend these surface criteria: it is not what is being said as what is *done* by the saying (Austin, 1962; Searle, 1975).

Category	*Example*
Questions	
Verification	Does the pumpkin land in his hands?
Disjunctive	Is the pumpkin accelerating or decelerating?
Concept Completion	Where will the pumpkin land?
Feature Specification	What are the components of the forces acting on the pumpkin?
Quantification	How far will the pumpkin travel?
Definition	What is acceleration?
Example	What is an example of Newton's Third Law?
Comparison	What is the difference between speed and velocity?
Interpretation	What is happening in this situation with the runner and pumpkin?
Causal Antecedent	What caused the pumpkin to fall?
Causal Consequence	What happens when the runner speeds up?
Goal Orientation	Why did you ignore air resistance?
Instrumental/Procedural	How do you calculate force?
Enablement	What principle allows you to ignore the vertical component of the force?
Expectational	Why doesn't the pumpkin land behind the runner?
Judgmental	What do you think of my explanation?
Frozen Expressions	
Metacognitive	I don't understand.
Metacommunicative	Could you repeat that?
Contribution	The pumpkin will land in the runner's hands

Table 1. AutoTutor's utterance taxonomy.

The close relation to speech acts underscores what a difficult task classifying conceptual questions can be. Jurafsky and Martin (2000) describe the problem of interpreting speech acts using pragmatic and semantic inference as AI-complete, i.e. impossible without creating a full artificial intelligence. The alternative explored in this paper is cue or surface-based classification, using no context.

It is particularly pertinent to the present discussion that the sixteen qualitative categories are employed in a quantitative classification process. That is to say that for the present purposes of classification, a question must belong to one and only one category. On the one hand this idealization is necessary to obtain easily analyzed performance data and to create a well-balanced training corpus. On the other hand, it is not entirely accurate because some questions may be assigned to multiple categories, suggesting a polythetic coding scheme (Graesser et al. 1992). Inter-rater reliability is used in the current study as a benchmark to gauge this potential effect.

Frozen expressions consist of metacognitive and metacommunicative utterances. Metacognitive utterances describe the cognitive state of the student, and they therefore require a different response than questions or assertions. AutoTutor responds to metacognitive utterances with canned expressions such as, "Why don't you give me what you know, and we'll take it from there." Metacommunicative acts likewise refer to the dialogue between tutor and student, often calling for a repetition of the tutor's last utterance. Two key points are worth noting: frozen expressions have a much smaller variability than questions or contributions, and frozen expressions may be followed by some content, making them more properly treated as questions. For example, "I don't understand" is frozen, but "I don't understand gravity" is a more appropriately a question.

Contributions in the taxonomy can be viewed as anything that is not frozen or a question; in fact, that is essentially how the classifier works. Contributions in AutoTutor, either as responses to questions or unprompted, are tracked to evaluate student performance via LSA, forming the basis for feedback.

4 Classifier Algorithm

The present approach ignores the semantic and pragmatic context of the questions, and utilizes surface features to classify questions. This shallow approach parallels work in question answering (Srihari and Li 2000; Soubbotin and Soubbotin 2002; Moldovan et al 1999). Specifically, the classifier uses tagging provided by ApplePie (Sekine and Grishman 1995) followed by cascaded finite state transducers defining the categories. The finite state transducers are roughly described in Table 2. Every transducer is given a chance to match, and a disambiguation routine is applied at the end to select a single category.

Immediately after tagging, transducers are applied to check for frozen expressions. A frozen expression must match, and the utterance must be free of any nouns, i.e. not frozen+content, for the utterance to be classified as frozen. Next the utterance is checked for question stems, e.g. WHAT, HOW, WHY, etc. and question mark punctuation. If question stems are buried in the utterance, e.g. "I don't know what gravity is", a movement rule transforms the utterance, placing the stem at the beginning. Likewise if a question ends with a question mark but has no stem, an AUX stem is placed at the beginning of the utterance. In this way the same transducers can be applied to both direct and indirect questions. At this stage, if the utterance does not possess a question stem and is not followed by a question mark, the utterance is classified as a contribution.

Two sets of finite state transducers are applied to potential questions, keyword transducers and syntactic pattern transducers. Keyword transducers replace a set of keywords specific to a category with a symbol for that category. This extra step simplifies the syntactic pattern transducers that look for the category symbol in their pattern. The definition keyword transducer, for example, replaces "definition", "define", "meaning", "means", and "understanding" with "KEYDEF". For most categories, the keyword list is quite extensive and exceeds the space limitations of Table 2. Keyword transducers also add the category symbol to a list when they match; this list is used for disambiguation. Syntactic pattern transducers likewise match, putting a category symbol on a separate disambiguation list.

In the disambiguation routine, both lists are consulted, and the first category symbol found on both lists determines the classification of the utterance. Clearly

<table>
<tr><th><em>Utterance Category</em></th><th><em>Finite state transducer pattern</em></th></tr>
<tr><td>Verification</td><td>^AUX</td></tr>
<tr><td>Disjunctive</td><td>^AUX ... or</td></tr>
<tr><td>Concept Completion</td><td>^(Who|What|When|Where)</td></tr>
<tr><td>Feature Specification</td><td>^What ... keyword
keyword</td></tr>
<tr><td>Quantification</td><td>^What AUX ... keyword
^How (ADJ|ADV)
^MODAL you ... keyword</td></tr>
<tr><td>Definition</td><td>^What AUX ... (keyword|a? (ADJ|ADV)* N
^MODAL you ... keyword
what a? (ADJ|ADV)* N BE</td></tr>
<tr><td>Example</td><td>^AUX ... keyword
^What AUX ... keyword</td></tr>
<tr><td>Comparison</td><td>^What AUX ... keyword
^How ... keyword
^MODAL you ... keyword</td></tr>
<tr><td>Interpretation</td><td>keyword</td></tr>
<tr><td>Causal Antecedent</td><td>^(Why|How) AUX ... (VBpast|keyword)
^(WH|How) ... keyword</td></tr>
<tr><td>Causal Consequence</td><td></td></tr>
<tr><td>Goal Orientation</td><td>^(What|Why) AUX ART? (NP|SUBJPRO|keyword)
^What ... keyword</td></tr>
<tr><td>Instrumental/Procedural</td><td>^(WH|How) AUX ART? (N|PRO)
^(WH|How) ... keyword
^MODAL you ... keyword</td></tr>
<tr><td>Enablement</td><td>^(WH|How) ... keyword</td></tr>
<tr><td>Expectational</td><td>^Why AUX ... NEG</td></tr>
<tr><td>Judgmental</td><td>(should|keyword) (N|PRO)</td></tr>
<tr><td>(you|your) ... keyword</td><td></td></tr>
<tr><td>Frozen (no nouns)</td><td>^SUBJPRO ... keyword
^VB ... keyword ... OBJPRO
^AUX ... SUBJPRO ... keyword</td></tr>
<tr><td>Contribution</td><td>Everything else</td></tr>
</table>

Table 2. Finite state transducer patterns

ordering of transducers affects which symbols are closest to the beginning of the list. Ordering is particularly relevant when considering categories like concept completion, which match more freely than other categories. Ordering gives rarer and stricter categories a chance to match first; this strategy is common in stemming (Paice 1990).

5 Training

The classifier was built by hand in a cyclical process of inspecting questions, inducing rules, and testing the results. The training data was derived from brainstorming sessions whose goal was to generate questions as lexically and syntactically distinct as possible. Of the brainstormed questions, only when all five raters agreed on the category was a question used for training; this approach filtered out polythetic questions and left only archetypes.

Intuitive analysis suggested that the majority of questions have at most a two-part pattern consisting of a syntactic template and/or a keyword identifiable for that category. A trivial example is disjunction, whose syntactic template is auxiliary-initial and corresponding keyword is “or”. Other categories were similarly defined either by one or more patterns of initial constituents, or a keyword, or both. To promote generalizability, extra care was given not to overfit the training data. Specifically, keywords or syntactic patterns were only used to define categories when they occurred more than once or were judged highly diagnostic.

	Expert	
Classifier	present	¬present
present	tp	fp
¬present	fn	tn

Table 3. Contingency Table.

The results of the training process are shown in Table 4. Results from each category were compiled in 2 x 2 contingency tables like Table 3, where *tp* stands for "true positive" and *fn* for "false negative".

Recall, fallout, precision, and f-measure were calculated in the following way for each category:

$$\text{Recall} = tp / (tp + fn)$$
$$\text{Fallout} = fp / (fp + tn)$$
$$\text{Precision} = tp / (tp + fp)$$

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Recall and fallout are often used in signal detection analysis to calculate a measure called d’ (Green and Swets 1966). Under this analysis, the performance of the classifier is significantly more favorable than under the F-measure, principally because the fallout, or false alarm rate, is so low. Both in training and evaluation, however, the data violate assumptions of normality that d’ requires.

As explained in Section 3, a contribution classification is the default when no other classification can be given. As such, no training data was created for contributions. Likewise frozen expressions were judged to be essentially a closed class of phrases and do not require training. Absence of training results for these categories is represented by double stars in Table 4.

During the training process, the classifier was never tested on unseen data. A number of factors it difficult to obtain questions suitable for testing purposes. Brainstormed questions are an unreliable source of testing data because they are not randomly sampled. In general, corpora proved to be an unsatisfactory source of questions due to low inter-rater reliability and skewed distribution of categories.

Low inter-rater reliability often could be traced to anaphora and pragmatic context. For example, the question "Do you know what the concept of group cell is?" might license a definition or verification, depending on the common ground. "Do you know what it is?" could equally license a number of categories, depending on the referent of "it". Such questions are clearly beyond the scope of a classifier that does not use context.

The skewed distribution of the question categories and their infrequency necessitates use of an extraction algorithm to locate them. Simply looking for question marks is not enough: our estimates predict that raters would need to classify more than 5,000 questions extracted from the Wall Street Journal this way to get a mere 20 instances of the rarest types. A bootstrapping approach using machine learning is a possible alternative that will be explored in the future (Abney 2002).

Regardless of these difficulties, the strongest evaluation results from using the classifier in a real world task, with real world data.

6 Evaluation

The classifier was used in AutoTutor sessions throughout the year of 2002. The log files from these sessions contained 9094 student utterances, each of which was classified by an expert. The expert ratings were compared to the classifier's ratings, forming a 2 x 2 contingency table for each category as in Table 4.

To expedite ratings, utterances extracted from the log files were split into two groups, contributions and non-contributions, according to their logged classification. Expert judges were assigned to a group and instructed to classify a set of utterances to one of the 18 categories. Though inter-rater reliability using the

	Training Data				*AutoTutor Performance*				
CATEGORY	*Recall*	*Fallout*	*Precision*	*F-measure*	*Recall*	*Fallout*	*Precision*	*F-measure*	*Likelihood Ratio*
Contribution	**	**	**	**	0.983	0.054	0.999	0.991	1508.260
Frozen	**	**	**	**	0.899	0.002	0.849	0.873	978.810
Concept Completion	0.844	0.035	0.761	0.800	0.857	0.003	0.444	0.585	235.800
Interpretation	0.545	0.009	0.545	0.545	0.550	0.000	0.917	0.688	135.360
Definition	0.667	0.002	0.941	0.780	0.424	0.001	0.583	0.491	131.770
Verification	0.969	0.004	0.969	0.969	0.520	0.004	0.255	0.342	103.880
Comparison	0.955	0.011	0.778	0.857	1.000	0.004	0.132	0.233	55.460
Quantification	0.949	0.002	0.982	0.966	0.556	0.003	0.139	0.222	43.710
Expecational	0.833	0.010	0.833	0.833	1.000	0.000	0.667	0.800	33.870
Procedural	0.545	0.009	0.545	0.545	1.000	0.000	1.000	1.000	20.230
Goal Orientation	0.926	0.006	0.893	0.909	1.000	0.001	0.143	0.250	14.490
Judgmental	0.842	0.010	0.865	0.853	0.500	0.001	0.167	0.250	12.050
Disjunction	0.926	0.000	1.000	0.962	0.333	0.000	0.250	0.286	11.910
Causal Antecedent	0.667	0.017	0.667	0.667	0.200	0.001	0.083	0.118	8.350*
Feature Specification	0.824	0.006	0.824	0.824	0.000	0.000	0.000	0.000	0.000*
Enablement	0.875	0.006	0.903	0.889	0.000	0.000	0.000	0.000	0.000*
Causal Consequent	0.811	0.008	0.882	0.845	0.000	0.000	0.000	0.000	0.000*
Example	0.950	0.008	0.826	0.884	**	**	**	**	**

Table 4. Training data and AutoTutor results.

kappa statistic (Carletta 1996) may be calculated for each group, the distribution of categories in the contribution group was highly skewed and warrants further discussion.

Skewed categories bias the kappa statistic to low values even when the proportion of rater agreement is very high (Feinstein and Cicchetti 1990a; Feinstein and Cicchetti 1990b). In the contribution group, judges can expect to see mostly one category, contribution, whereas judges in the non-contribution group can expect to see the other 17 categories. Expected agreement by chance for the contribution group was 98%. Correspondingly, inter-rater reliability using the kappa statistic was low for the contribution group, .5 despite 99% proportion agreement, and high for non-contribution group, .93.

However, the .93 inter-rater agreement can be extended to all of the utterance categories. Due to classifier error, the non-contribution group consisted of 38% contributions. Thus the .93 agreement applies to contributions in this group. Equal proportion of agreement for contribution classifications in both groups, 99%, suggests that the differences in kappa solely reflect differences in category skew across groups. Under this analysis, dividing the utterances into two groups improved the distribution of categories for the calculation of kappa (Feinstein and Cicchetti 1990b).

Expert judges classified questions with a .93 kappa, which supports a monothetic classification scheme for this application. In Section 3 the possibility was raised of a polythetic scheme for question classification, i.e. one in which two categories could be assigned to a given question. If a polythetic scheme were truly necessary, one would expect inter-rater reliability to suffer in a monothetic classification task. High inter-rater reliability on the monothetic classification task renders polythetic schemes superfluous for this application.

The recall column for evaluation in Table 4 is generally much higher than corresponding cells in the precision column. The disparity implies a high rate of false positives for each of the categories. One possible explanation is the reconstruction algorithm applied during classification. It was observed that, particularly in the language of physics, student used question stems in utterances that were not questions, e.g. "The ball will land when …" Such falsely reconstructed questions account for 40% of the questions detected by the classifier. Whether modifying the reconstruction algorithm would improve F-measure, i.e. improve precision without sacrificing recall, is a question for future research.

The distribution of categories is highly skewed: 97% of the utterances were contributions, and example questions never occurred at all. In addition to recall, fallout, precision, and F-measure, significance tests were calculated for each category's contingency table to insure that the cells were statistically significant. Since most of the categories had at least one cell with an expected value of less than 1, Fisher's exact test is more appropriate for significance testing than likelihood ratios or chi-square (Pedersen 1996). Those categories that are not significant are starred; all other categories are significant, $p < .001$.

Though not appropriate for hypothesis testing in this instance, likelihood ratios provide a comparison of classifier performance across categories. Likelihood ratios are particularly useful when comparing common and rare events (Dunning 1993; Plaunt and Norgard 1998), making them natural here given the rareness of most question categories and the frequency of contributions. The likelihood ratios in the rightmost column of Table 4 are on a natural logarithmic scale, $-2\ln\lambda$, so procedural at $\boldsymbol{e}^{.5\ \mathrm{x}\ 20.23} = 24711$ is more likely than goal orientation, at $\boldsymbol{e}^{.5\ \mathrm{x}\ 14.49} = 1401$, with respect to the base rate, or null hypothesis.

To judge overall performance on the AutoTutor sessions, an average weighted F-measure may be calculated by summing the products of all category F-measures with their frequencies:

$$F_{avg} = \sum F - measure \times \frac{tp + fn}{N}$$

The average weighted F-measure reflects real world performance since accuracy on frequently occurring classes is weighted more. The average weighted F-measure for the evaluation data is .98, mostly due to the great frequency of contributions (.97 of all utterances) and the high associated F-measure. Without weighting, the average F-measure for the significant cells is .54.

With respect to the three applications mentioned, i) tracking student understanding, ii) mixed-initiative dialogue, and iii) questions answering, the classifier is doing extremely well on the first two and adequately on the last. The first two applications for the most part require distinguishing questions from contributions, which the classifier does extremely well, F-measure = .99. Question answering, on the other hand, can benefit from more precise identification of the question type, and the average unweighted F-measure for the significant questions is .48.

7 Conclusion

One of the objectives of this work was to see how well a classifier could perform with a minimum of resources. Using no context and only surface features, the classifier performed with an average weighted F-measure of .98 on real world data.

However, the question remains how performance will fare as rare questions become more frequent. Scaffolding student questions has become a hot topic recently (Graesser et al. 2003). In a system that greatly promotes question-asking, the weighted average of .97 will tend to drift closer to the unweighted average of .54. Thus there is clearly more work to be done.

Future directions include using bootstrapping methods and statistical techniques on tutoring corpora and using context to disambiguate question classification.

8 Acknowledgements

This research was supported by the Office of Naval Research (N00014-00-1-0600) and the National Science Foundation (SBR 9720314 and REC 0106965). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

References

Abney, Steven. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 360-367.

Allen, J.F. 1999. Mixed Initiative Interaction. *Proc. IEEE Intelligent Systems 14*(6).

Austin, John. 1962. *How to do things with words*. Harvard University Press, Cambridge, MA.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics, 21*(4), 543-566.

Brill, Eric, J. Lin, M. Banko, S. Dumais, and A. Ng. 2001. Data-intensive question answering. *Proceedings of the 10th Annual Text Retrieval Conference (TREC-10).*

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22*(2), 249-254.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*, 61-74.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990a. High agreement but low kappa: the problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543-549.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990b. High agreement but low kappa: II. resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551-558.

Graesser, Arthur, John Burger, Jack Carroll, Albert Corbett, Lisa Ferro, Douglas Gordon, Warren Greiff, Sanda Harabagiu, Kay Howell, Henry Kelly, Diane Litman, Max Louwerse, Allison Moore, Adrian Pell, John Prange, Ellen Voorhees, and Wayne Ward. 2003. Question generation and answering systems, R&D for technology-enabled learning systems: research roadmap. Unpublished manuscript.

Graesser, Arthur, Natalie Person, and John Huber. 1992. Mechanisms that generate questions. In T. Lauer, E. Peacock, and A. Graesser (Eds), *Questions and information systems*. Earlbaum, Hillsdale, NJ.

Graesser, Arthur and Natalie Person. 1994. Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104-137.

Graesser, Arthur, Natalie Person, and J.P. Magliano. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359-387.

Graesser, Arthur, Kurt van Lehn, Carolyn Rose, Pamela Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine 22*(4), 39-52.

Graesser, Arthur, Peter Wiemer-Hastings, K. Wiemer-Hastings, Roger Kreuz, and the TRG. 1999. AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research 1*, 35-51.

Green, David and John Swets. 1966. *Signal detection theory and psychophysics.* John Wiley, New York.

Harabagiu, Sanda, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu. 2000. FALCON: Boosting knowledge for answer engines. In *Proceedings of the 9th Text Retrieval Conference* (TREC-9).

Jurafsky, Daniel and James Martin. 2000. *Speech and language processing*. Prentice Hall, NJ.

Lehnert, Wendy. 1978. *The Process of Question Answering*. Lawrence Erlbaum Associates, Hillsdale, NJ.

McArthur, D., C. Stasz, and M. Zmuidzinas. 1990. Tutoring techniques in algebra. *Cognition and Instruction*, *7*, 197-244.

Moldovan, Dan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Girju, and Vaslie Rus. 1999. Lasso: a tool for surfing the answer net *Proceedings of the 8th Annual Text Retrieval Conference (TREC-8)*, 65-73.

Olney, Andrew, Natalie Person, Max Louwerse, and Arthur Graesser. 2002. AutoTutor: a conversational tutoring environment. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Demonstration Abstracts*, 108-109.

Otero, J. and Arthur Graesser. 2001. PREG: Elements of a model of question asking. *Cognition & Instruction 19*, 143-175.

Paice, C.D. 1990. Another stemmer. *SIGIR Forum 24* (3), 56-61.

Pedersen, Ted. 1996. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX.

Person, Natalie and Arthur Graesser. 1999. Evolution of discourse in cross-age tutoring. In A.M. O'Donnell and A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 69-86). Erlbaum, Mahwah, NJ.

Person, Natalie, Arthur Graesser, L. Bautista, E.C. Mathews, and the Tutoring Research Group 2001. Evaluating student learning gains in two versions of AutoTutor. In J. D. Moore, C. L. Redfield, and W. L. Johnson (Eds.) *Artificial intelligence in education: AI-ED in the wired and wireless future* (pp. 286-293). IOS Press, Amsterdam.

Person, Natalie, Arthur Graesser, Derek Harter, E. C. Mathews, and the Tutoring Research Group (2000). Dialog move generation and conversation management in AutoTutor. *Proceedings for the AAAI Fall Symposium Series: Building Dialogue Systems for Tutorial Applications*. Falmouth, Massachusetts.

Plaunt, Christian and Barbara Norgard. 1998. An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society of Information Science, 49*(10), 888-902.

Putnam, R. T. 1987. Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24,* 13-48.

Searle, John. 1975. A taxonomy of illocutionary acts. In K. Gunderson, (Ed.), *Language, mind, and knowledge*. University of Minnesota Press, Minneapolis, MN.

Sekine, S. and R. Grishman. 1995. A corpus-based probabilistic grammar with only two nonterminals. *Fourth International Workshop on Parsing Technology*.

Soubbotin, M. M., and S. M. Soubbotin. 2002. Patterns of potential answer expressions as clues to the right answers. *Proceedings of the 10th Annual Text Retrieval Conference (TREC-10).*

Srihari, Rohini and Wei Li. 2000. A question answering system supported by information extraction. *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-2000)*, 166-172.

Van Dijk, T. A., and W. Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic.

Voorhees, Ellen. 2001. Overview of the TREC 2001 question answering track. *Proceedings of the 10th Annual Text Retrieval Conference (TREC-10)*, 400-410.