On the Types, Frequency, Uses and Characteristics of Meta-language in Conversation

Michael L. Anderson (anderson@cs.umd.edu)
Institute for Advanced Computer Studies
University of Maryland; College Park, MD 20742 USA

Bryant Lee, Jon Go, Shuda Li, Ben Sutandio and LuoYan Zhou
University of Maryland
College Park, MD 20742 USA

Abstract

Human dialog is a highly collaborative and interactive process, that includes the ability to talk about the dialog itself and its linguistic constituents, and to use meta-linguistic interactions to help coordinate the ongoing conversation. However, very little is known about the frequency and conditions under which people resort to meta-language, its range of uses, and any characteristics that may be useful to its automated identification. This paper presents the results of a corpus study in which a markup scheme for meta-language was applied to a sub-set of the British National Corpus. The corpus study made it possible to demonstrate that sentences containing meta-language show a high degree of correlation with instances of dialog management, and that automated detection of meta-language should be feasible, based on word-frequency analysis.

Introduction and Background

It is well-established that meta-linguistic skills play an important role in language learning, for instance in talking about the meanings, spellings, and proper use of words [Clark and Schaefer, 1989]. Researchers also believe that the ability to engage in meta-language is necessary for the adult ability to engage in free and flexible conversation, and more importantly, that a robust meta-dialogic ability can make up for weaknesses in other areas of linguistic ability [Perlis et al., 1998]. For this reason, we think that time spent understanding meta-language, and implementing meta-linguistic skills in natural-language HCI systems, will be well rewarded; the ability to engage in even simple meta-language can be used to fruitfully enhance the performance of interactive systems, even those having relatively limited speech-recognition and language-processing abilities. The work described here is part of a larger project involving the development of viable natural-language computer interfaces with the ability to engage in meta-language, and thereby with some of the flexibility that meta-language provides to human conversation.

Natural language is complex and ambiguous, and communication for this reason always contains an element of uncertainty. To manage this uncertainty, dialog partners continually monitor their conversations, their own comprehension, and the apparent comprehension of their interlocutor, routinely eliciting and providing feedback as the conversation continues [Brennan, 1998, Brennan, 2000, Brennan and Hulteen, 1995, Cahn and Brennan, 1999, Clark and Schaefer, 1987, Clark and Schaefer, 1989, Clark and Brennan, 1991, Krahmer et al., 2001, Paek and Horvitz, 1999, Traum, 1994]. The feedback might be as simple as “Got it?”, eliciting a simple “yes”, or as complex as “Wait. I don’t think I understand the concept of hidden variables”, which could result in a long digression.

Dialog annotation schemes generally recognize this fact by providing markers for utterances aimed at task and conversation management, as well as for such things as overtures and acceptances (see e.g. [Allen and Core, 1997]). There also exist annotation schemes specifically for dialog clarifications [Purver, 2002, Purver et al., 2002, Heeman and Allen, 1994], as well as schemes for annotating self-correction in spoken dialog [Bear et al., 1992, Core and Schubert, 1998, Heeman and Allen, 1994]. However, there are currently no schemes or studies which focus on meta-language in particular, nor on the full range of meta-linguistic behavior in conversation. Given the apparent importance of meta-language to human conversation, and the hypothesis that conversational adequacy requires facility with meta-reasoning and meta-language [Anderson et al., 2002, Anderson et al., 2003, Perlis et al., 1998], we have begun to address this lacuna.

Three Studies of Meta-language

In this paper we report the results of three related studies. The first was the development of an annotation scheme for meta-language, which was then applied to a 59 file sub-set of the British National Corpus (BNC). The second was a small study that used the results of this annotation for three selected BNC files, and correlated meta-language with instances of dialog management, producing results that confirmed the hypothesis that meta-language is frequently involved in dialog management. The third study analyzed the differences between sentences that contain meta-language and sentences that do not, with respect to vocabulary and word frequency. The results found many significant differences, suggesting the feasibility of using simple naive Bayes methods of document classification [McCallum and Nigam, 1998] to automatically detect and classify sentences containing meta-language.

The Development and Application of an Annotation Scheme for Meta-language

Because meta-language is often used for dialog clarification or repairs, we began with a study of currently existing schemes for classifying those speech acts [Purver, 2002, Purver et al., 2002, Heeman and Allen, 1994, Allen and Core, 1997]. However, it quickly became clear that not all meta-language is a repair, nor do all repairs involve meta-language. Thus, our first task was to re-define and greatly expand the scope of the annotations to include the many types of meta-language that are not, in fact, repairs. To maximize comparability with existing annotations for clarification and repairs, we decided to study the very same 59 file sub-set of the BNC used to develop the annotations described in [Purver, 2002, Purver et al., 2002].[1] In addition, the use of the BNC, which is a repository of general conversations and dialogs, to develop the annotation scheme helps to limit any bias that could result from the use of a more narrowly focused or specialized dialog corpus.

We used a 3-step approach. First, each BNC file in the sub-set was assigned to two different researchers. The researchers consisted of four undergraduate research assistants and one faculty member. Each researcher separately read the files and identified possible instances of meta-language. To guide the identification, the definition of meta-language used was: language that refers or explicitly draws attention to speech acts, to items in the language, or to aspects of these other than their extension. This definition is adopted from Paul Saka [Saka, 1998], and was chosen because it best reflects the breadth of the phenomenon under consideration [Anderson et al., 2002]. The union of the different annotations were then read by two different researchers, who confirmed or rejected each item as an instance of meta-language. At the end of this process we were left with a set of files from the identified sub-set of the BNC, with meta-linguistic dialog exchanges marked-up.

As our main interest was ensuring we gathered all the relevant examples (and did not include false positives), we did not keep data for the degree of inter-annotator overlap at this stage. However, the *rejection rate*–the amount of meta-language that was determined to be incorrectly identified–was very low, less than 1%. So, while we have no data about how easy it is to find *all* instances of meta-language (the false negative rate), the true postive rate is very high.

Next, we used this set of meta-language files to develop an annotation scheme. The development process worked in the following way: a preliminary scheme was proposed, and at least two researchers separately attempted to apply this scheme to a few BNC meta-language files. The results were then evaluated with respect to their coverage (the number of instances of meta-language that fell into one of the categories) and their reliability (the amount of agreement between the two researchers). Instances of conflict, as well as confusing and difficult cases were discussed, and modifications to the scheme were proposed in light of this discussion. The end result of this iterative process was a pragmatic annotation scheme with five major categories:

Track Dialog (TD) Interchanges used to establish, track, and move between dialog states. For instance,

– "Which particular section of the conversation are we talking about?" [BNC KPK.860]

Speaker Meaning (SM) Interchanges used to establish communicative intention, or speaker meaning. For instance,

– "I had a right argument over that"
– "Who did, them two?"
– "No, me and Laura did." [BNC KSW.1008-10]

Language Meaning (LM) Discussions about or clarifications of items in the language itself (e.g. parts of speech, spelling, word meanings, etc.). For instance,

– "So you have a bilge, and, you eat loads of cakes and then instead of like you with pizzas down there, they just throw it up."
– "Yes, as well, binge, binge, not 'bilge'." [BNC KPL.547-8]

Determine Truth (DT) Interchanges used to establish or monitor the match between language and the world. For instance,

– "I'd rather be working."
– "Oh, God. You don't really mean it?" [BNC KSU.460-3]

Speech Acts (SA) Discussions of or references to speech acts *per se*, including such things as their content, timing, style, appropriateness, effectiveness, etc. For instance,

– "Yeah, we remember when you shouted 'here she comes'." [BNC KSW.814]

Quotation belongs in this category.

Third and finally, we applied the annotation scheme to the entire 59-file sub-set of the BNC. Here again, two different researchers separately applied the annotations to each file, and the results were compared. The reliability of the annotation scheme was a measure of the agreement between the different researchers in the application of the annotation scheme.

After the reliability of the scheme was determined, instances of conflict were discussed, and if consensus could

[1] Note, however, that Purver, et al. annotated only 200 speaker turns from each file, covering a total of 18,983 lines, whereas we annotated the entire file. See results, below.

Type	Number	Percentage
TD	1185	7.66%
SM	2340	15.13%
LM	593	3.83%
DT	170	1.10%
SA	11154	72.10%
O	28	0.18%

Table 1: Distribution of meta-language types

be reached on how the instance should be classified, it was placed in that category. Items which could not be classified, or about which no agreement could be reached, were classified as "Other".

Results from the Annotation Study Overall, a total of 137,921 sentences from the BNC were examined. Of these, 15,091 sentences (10.94%) were identified as containing meta-language. In some cases a single sentence contained more than one instance of meta-language, and so the total number of *instances* of meta-language is 15,470. The instances of meta-language break down into categories as shown in Table 1. The table shows each type of meta-language, and the number of lines containing an instance of that type.

By way of comparison, Purver et al. [Purver, 2002, Purver et al., 2002] studied 200-turn extracts from each of the 59 files, covering 18,983 lines, and found 418 clarification requests. In the same range, we discovered 1,924 instances of meta-language, 165 of which were counted by Purver et al. as clarification requests. 152 of these were categorized by us as SM, 7 as SA, 3 as DT, 2 as LM and 1 as TD. This leaves 253 clarification requests not counted as meta-language, and 1759 instances of meta-language not counted as clarification requests. This latter figure underscores the broader scope of our annotation scheme for meta-language.

The reason for the high number of clarification requests that were not considered meta-language is that we did not consider most reprise clarifications to be meta-linguistic in form. For instance,

– "I spoke to him on Wednesday, I phoned him."
– "You phoned him?"
– "Phoned him." [BNC KPW.463-5]

counts as a literal reprise clarification for Purver et al., but is not meta-linguistic in form.

The evaluation criteria most important to the annotation scheme are coverage and reliability. Reliability results are determined by comparing the annotation results of the same set of sentences by two or more annotators, and determining the percentage of agreement in the different annotations.

Coverage results are calculated after any conflicts revealed while evaluating the annotations for reliability are discussed, and, where possible, adjudicated. Instances of meta-language which cannot be fit into any category, or on which no agreement as to its category can be reached, are labeled "Other". Coverage is a measure of the percentage of instances of meta-language which are not labeled "Other". The reliability of this scheme was 95%, and its coverage was $>$ 99%.

While developing the above annotation scheme, we also began to develop a set of sub-categories for each of these major categories; the category SA in particular is in need of sub-division, to account for differences between direct and indirect quotation, as well as uses of meta-language referring to non-content-related characteristics of the speech act, such as timing, style, appropriateness, etc. At present, however, the sub-categories provide less coverage and reliability than is necessary for a maximally useful annotation scheme. Thus, among our future tasks will be improving the coverage and reliability of the sub-categories.

Meta-language and Dialog Management

A second study, in which we put to use the above annotations, investigated the correlation between meta-language and dialog-management utterances in three dialog files of the BNC.

The three BNC files used in this study, KRF, KRG, and KRH, are transcripts of a series of Ideas in Action radio programs, some of which are interviews. Because interviews are more structured than informal conversation, they involve explicit dialog management, and are therefore a good place to start an investigation into the relation between meta-language and dialog management. Focusing exclusively on the interviews in these three files gives 5900 lines to study. Naturally, these three files had been previously annotated for meta-language, using the annotation scheme and methods described above.

Two different counting methods for dialog management utterances were used. Both were developed using Dialog Act Markup in Several Layers (DAMSL) [Allen and Core, 1997], a method for identifying and tagging speech acts in task-oriented dialog. DAMSL identifies three different information levels: task, task management, and communication management. The task level encompasses utterances directly involved in "performing the task that is the reason for the dialog" (tasks are generally imagined to be such collaborative endeavors as fixing a car) and utterances that "directly move ahead (or attempt to move ahead) the goals of the domain." The task management level, in contrast, "explicitly addresses the problem solving process", and "includes utterances that involve coordinating the activities of the two speakers, asking for help on the procedures, or asking about the status of the process." Finally, the communication management level includes "conventional phrases that maintain contact, perception, and understanding during the communication process."

For the first counting method, a very strict interpretation of DAMSL was used, wherein the task was defined as performing/participating in an interview, and strict interpretations of each level were used. Thus, for instance, on this interpretation task level utterances include discussing the interview topic, talking about what was said (e.g. "you said 'x' before") including summarizing, clarifying utterances, requesting utterances, and the like. In contrast, task management utterances would include such things as agreeing on a topic of conversation, changing the topic of conversation, requesting permission to talk about a subject, talking about the format of the interview, etc. The advantage of this method of counting is that it is in strict adherence with a well-established method, allowing a high degree of confidence in the results. However, the disadvantage, as may be apparent from the above list, is that many things that qualify as task level on a strict interpretation of DAMSL, probably in fact belong in the category of dialog management, such as clarifying and requesting utterances.

Thus, the second counting method involved modifying DAMSL somewhat to better fit the case where the task under consideration is an interview. When the task is itself a discussion, two issues need to be addressed. First, the distinction between the last two information levels blurs somewhat; task management can be considered a kind of communication management. Second, as mentioned above, some task level utterances, that do not involve discussion of the dialog *per se*, are effectively part of the dialog management. An example of such an utterance is: "And can you give me some examples of the firms that the University's managed to help?" (BNC KRH 818).

To address these issues, we (a) collapsed the task management and communication management levels, categorizing all relevant utterances under the latter, and (b) added a dialog-management marker, applied on top of the standard markers, for utterances having an explicit, intended effect on the course of the discussion. Such utterances, along with the communication-management utterances, were counted as instances of dialog management. The advantage of this second counting method is that we can be more certain that all the dialog management has been counted. The disadvantage is that, since the method is new, it is not known how reliable it is. However, using the two methods together, we can be fairly certain that the overall results are sound.

Results from Dialog Management Study Of the 5900 lines annotated, 1086 included meta-language. Given that only 18.4% of the annotated lines included meta-language, if there were no relation between meta-language and dialog management, then we would expect only 18.4% of dialog-management lines to also contain meta-language. Any significant deviation from this distribution would indicate a correlation.

Results from Method 1: Of the 5900 lines annotated, there were 270 task-management utterances. Thus, the null hypothesis would predict that 18.4% of these, or about 50 lines, would also contain meta-language. However, in fact 151 lines were both dialog management and meta-language, giving $\chi^2 = 265.2$, $p \ll .001$, and $\Phi = 0.212$. By this counting method, 60.37% of dialog-management utterances involved meta-language.

Results from Method 2: In the 5900 lines annotated, there were 741 dialog-management utterances. Thus, the null hypothesis would predict that 18.4% of these, or about 136 lines, would also contain meta-language. However, in fact 407 lines were both dialog management and meta-language, giving $\chi^2 = 753.74$, $p \ll .001$, and $\Phi = 0.357$. By this counting method, 54.93% of dialog-management utterances involved meta-language.

Both methods gave results that were largely in accord, thus confirming the hypothesis that meta-language is frequently involved in dialog management. To our knowledge this is the first empirical confirmation of this presumably widely-shared assumption regarding the use of meta-language.

In the sample studied, there were 112 meta-language utterances annotated as **TD**. 97 of these (86.61%) were also tagged as dialog-management utterances (using the second counting method). The high correlation between meta-linguistic "track dialog" utterances and DAMSL dialog-management utterances is of course to be expected. More interesting is the fact that 310 *other* meta-linguistic utterances functioned as dialog-management. Of these, 267 were SA, indicating that making direct reference to speech acts is an important strategy in dialog-management.

Word-frequency Markers for Meta-language

For this study we looked for word frequency differences between sentences that contained meta-language, and those that did not. To do this, we first separated out the relevant sentences from the BNC, and created two different files, one containing the meta-language sentences, and one containing the non-meta-language sentences. The word count was done using PhraseContext [Mortensen, 2005], supplemented with software developed by the research group for counting the number of lines containing a given word.

Given that in the sub-set of the BNC studied there are 15,091 sentences that contain meta-language and 122,830 that do not, out of 137,091 total sentences, one would expect a given word to appear 10.94% of the time in meta-language sentences, and 89.06% of the time in non-meta-language sentences. Any significant deviation from this distribution was of interest. For the purposes of this study, we defined "significant" as any distribution where $\chi^2 > 10.83$, which makes $p < .001$.

Results from the Word-frequency Study Our initial analysis shows that there are many significant differences in word frequencies between sentences containing meta-language versus those that do not. In fact, of the top 1000 most common words in our sub-set of the BNC, over half show statistically significant deviations from expected distributions. Table 2 shows the top 10 words most correlated with meta-language.

TP stands for true positive, or the number of lines of meta-language in which the word appeared, FP stands for false positive, or the number of lines of non-meta-language in which the word appeared, and PPV stands for positive predictive value, where $PPV = TP/(TP + FP)$. PPV is a measure of the value of a test in a classification task; roughly speaking, if a given test is positive, PPV is the likelihood that the subject belongs in

Word	TP	FP	PPV
said	2387	470	0.84
pardon	130	39	0.77
say/s/ing	3636	1317	0.73
talk/ing	1025	463	0.69
ask/ed/ing	654	321	0.67
told	335	196	0.63
speak/ing	207	127	0.62
tell/ing	694	463	0.60
question/s	337	252	0.57
word/s	278	240	0.54

Table 2: Ten words correlated with meta-language

the class of interest. In our case, the test is the appearance of a word in a sentence, and the class of interest is sentences containing meta-language.

The large number of deviations from expected distributions—that is, the large number of words that can be used as a test to indicate that a sentence contains meta-language—strongly suggests that word-frequency information can indeed be used as the basis of an automated classification system (keeping in mind that even small deviations can support reliable classifications when properly summed). Such a system is a necessary first step toward building any natural-language HCI device with the ability to recognize and interpret meta-language.

Discussion

Taken together, these three studies demonstrate that meta-language is a genuine class of speech-act, that it plays an important role in dialog-management, and that it has characteristics at least theoretically detectable by automated means. At the very least, this should alert other researchers to the importance of meta-language, and the potential utility of its study. For instance, it is our own long-term goal to use these results to help in the development of natural-language HCI systems, with the ability to engage in meta-language, and thereby with some of the flexibility that meta-language provides to human conversation. However, there are doubtless other uses to which these results can be put.

On the other hand, the results also highlight some places where improvements are needed. For instance, in Table 2, above, note that 470 instances of "said" appeared in sentences classified as not containing meta-language. It is very hard to imagine many uses of "said" that would not be meta-linguistic in nature; therefore, this likely reveals the existence of some mistakes in the initial classification. Such mistakes will have to be discovered and corrected, a task that will actually be helped along by building an automated meta-language classifier.

Further, the very large number of instances of meta-language falling in the category SA indicate the immediate need for a set of sub-categories to better account for and classify at least these instances of meta-language. Devising such a set of sub-categories will be among the first of our tasks going forward.

Future Tasks

Following the same method as described in the section describing our approach, above, our first task will be to develop a reliable set of sub-categories for our existing annotation scheme for meta-language. Once a reliable set of sub-categories has been developed, we plan to apply the scheme to the entire Map-Task, TRAINS-91, and TRAINS-93 corpora. Here again, we will follow the same method as employed in the preliminary study. Note that since the annotation scheme is being developed on a general corpus, and applied to more specialized corpora, it is likely that there will be some difference in its coverage and reliability when applied to these latter corpora. We do not expect a large difference; however, if we record a significant drop in the measured quality of the annotation scheme, we will attempt to adjust the scheme appropriately, following the methods outlined above.

In addition to straightforward statistical studies to determine the frequency of various types of meta-language, we will also cross-index our findings with existing annotations of these corpora for larger-scale dialog structures, (e.g. dialog moves and/or speech acts) as well as local syntax. We will be looking for correlations which could be used to help automated dialog systems recognize and categorize instances of meta-language, and appropriately interpret them in light of the conversational and task contexts. Good models for this task include the recent work by Adrian Bangerter and Herbert Clark [Bangerter and Clark, 2003], which correlated instances of feedback words (e.g. uh-huh, m-hm, yeah, okay, all-right) with horizontal and vertical transitions in ongoing joint projects between the dialog parters, as well as our own preliminary study correlating meta-language with dialog-management utterances, reported above.

Meanwhile, starting with the results from our preliminary work with the BNC, and working in parallel with the ongoing annotations, we will begin work on an automated meta-language classifier, able to detect, and perhaps type instances of meta-language. This ability is a crucial first step to properly processing such sentences in any natural-language HCI system. For this application, will examine Naive Bayes methods [McCallum and Nigam, 1998], Support Vector Machines [Yang and Liu, 1999], Linear Least Squares Fit Mappings [Yang and Chute, 1992], and k-Nearest Neighbor Classifiers [Denoeux, 1995, Han et al., 2001, Yang and Liu, 1999]; however, preliminary investigations suggest that the kNN classifiers will prove to be the most efficient and effective system for our purposes.

Acknowledgments

This work was supported in part by the Office of Naval Research, grants N000140310049 and N000140610057.

References

Allen, J. and Core, M. (1997). DAMSL: Dialog Annotation Markup in Several Layers. Technical report, University of Rochester.

Anderson, M. L., Josyula, D., and Perlis, D. (2003). Talking to computers. In *Proceedings of the Workshop on Mixed Initiative Intelligent Systems, IJCAI-03*, pages 1–8.

Anderson, M. L., Okamoto, Y., Josyula, D., and Perlis, D. (2002). The use-mention distinction and its importance to HCI. In *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialog*, pages 21–28.

Bangerter, A. and Clark, H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, 27(2):195–225.

Bear, J., Dowding, J., and Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th annual meeting of the Association for Computational Linguistics*, pages 56–63.

Brennan, S. E. (1998). The grounding problem in conversations with and through computers. In Fussell, S. and Kreuz, R., editors, *Social and Cognitive Psychological Approaches to Interpersonal Communication*, pages 201–225. Lawrence Erlbaum.

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Lingusitics*, pages 1–11.

Brennan, S. E. and Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, 8:143–151.

Cahn, J. E. and Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. In *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 25–33.

Clark, H. and Brennan, S. E. (1991). Grounding in communication. In L.B. Resnik, J. L. and Teasley, S., editors, *Perspectives on Socially Shared Cognition*, pages 127–149.

Clark, H. and Schaefer, E. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2:19–41.

Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

Core, M. and Schubert, L. (1998). Implementing parser metarules that handle speech repairs and other disruptions. In *Proceedings of the 11th Annual International FLAIRS conference*, pages 283–288.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813.

Han, E.-H. S., Karypis, G., and Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification. *Lecture Notes in Computer Science*, 2035:53–65.

Heeman, P. A. and Allen, J. (1994). Tagging speech repairs. In *Proceedings of the ARPA workshop on human language technology*, pages 187–192.

Krahmer, E., Swerts, M., Theune, M., and Weegels, M. (2001). Error detection in spoken human-machine interaction. *International Journal of Speech Technology*, 4(1):19–30.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, pages 41–48.

Mortensen, H. J. K. (2005). Phrasecontext: textual analysis and writing tool. http://www.hjkm.dk/PhraseContext/.

Paek, T. and Horvitz, E. (1999). Uncertainty, utility and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *Proceedings, AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 85–92.

Perlis, D., Purang, K., and Andersen, C. (1998). Conversational adequacy: mistakes are the essence. *Int. J. Human-Computer Studies*, 48:553–575.

Purver, M. (2002). A clarification request markup scheme for the BNC. Technical Report TR-02-02, Department of Computer Science, King's College London.

Purver, M., Ginzburg, J., and Healey, P. (2002). On the means for clarification in dialogue. In Smith, R. and van Kuppvelt, J., editors, *Current and New Directions in Discourse and Dialogue*, pages 235–255. Kluwer Academic Publishers.

Saka, P. (1998). Quotation and the use-mention distinction. *Mind*, 107:113–35.

Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester.

Yang, Y. and Chute, C. G. (1992). A linear least squares fit mapping method for information retrieval from natural language texts. In *Proceedings of the 14th conference on Computational linguistics*, pages 447–453, Morristown, NJ, USA. Association for Computational Linguistics.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley.