On the types and frequency of meta-language in conversation: A preliminary report

Michael L. Anderson, **Andrew Fister**, **Bryant Lee**, **Luwito Tardia** and **Danny Wang**
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Abstract

Human dialog is a highly collaborative and interactive process, which 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, the frequency and conditions under which people resort to meta-language are not well known. 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.

Introduction and Background

We use the term *conversational adequacy* to denote the ability to engage in free and flexible conversation. It is our contention that the ability to engage in meta-language is necessary for conversational adequacy, and more importantly, that a robust meta-dialogic ability can make up for weaknesses in other areas of linguistic ability [18]. For this reason, we think that time spent understanding and implementing meta-language 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 [11,12,6,7,8,9,10,15,16,17,21]. Dialog annotation schemes generally recognize this fact by providing markers for utterances aimed at task and conversation management, as well as such things as overtures and acceptances (see e.g. [1]). There also exist annotation schemes specifically for dialog clarifications [19,20,14], as well as schemes for annotating self-correction in spoken dialog [5,13,14]. However, there are currently no schemes or studies that focus on meta-language in particular, nor on the full range of meta-lingustic behavior in conversation. Given the apparent importance of meta-language to human conversation, and our hypothesis that conversational adequacy requires facility with meta-reasoning and meta-language [3,2,18], we have begun to address this lacuna.

Approach and Methods

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 [19,20,14,1]. 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 the contrast with existing annotations for clarification and repairs, we decided to study the very same fifty-nine file sub-set of the British National Corpus (BNC) used to develop the annotations described in [19,20]. (See Appendix 1 for a list of the files used.) 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 of the fifty-nine BNC files 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, which were then copied with surrounding context into a separate file. The union of these files was 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 containing meta-linguistic dialog exchanges and their context from the identified sub-set of the BNC.

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:

(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]

(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]

(ML) 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]

(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]

(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 sub-set of the BNC. (See Appendix 1 for the list of files used.) 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 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

A total of fifty-nine files containing 138,017 sentences from the BNC were examined. Of these, 15,832 lines (11.47%) were identified as containing meta-language. Overall, the instances of meta-language break down into categories as shown in Table 1. The full, detailed annotation can be downloaded from http://www.cs.umd.edu/projects/metalanguage.

Type	Number	Percentage
TD	1190	0.86%
SM	2363	1.71%
ML	595	0.43%
DT	170	0.12%
SA	11486	8.32%
O	28	0.02%
Totals	**15,832**	**11.47%**

Table 1: Frequency results for different types of meta-language in a sub-set of the British National Corpus.

Evaluation

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, of 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 that 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; 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.

Future Tasks

Following the same method as described 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 this 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 that 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. One model for this task is the recent work by Adrian Bangerter and Herbert Clark [4], which correlated instances of feedback words (e.g. uh-huh, m-hm, yeah, okay, allright) with horizontal and vertical transitions in ongoing joint projects between the dialog parters.

We expect that such results will be extremely useful for natural language HCI system designers. It is a 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.

Acknowledgements

This work was supported in part by the ONR.

References

[1] James Allen and Mark Core. DAMSL: Dialog Annotation Markup in Several Layers. Technical report, University of Rochester, 1997.

[2] Michael L. Anderson, Darsana Josyula, and Don Perlis. Talking to computers. In *Proceedings of the Workshop on Mixed Initiative Intelligent Systems, IJCAI-03*, 2003.

[3] Michael L. Anderson, Yoshi Okamoto, Darsana Josyula, and Don Perlis. The use-mention distinction and its importance to HCI. In *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialog*, 2002.

[4] Adrian Bangerter and Herbert Clark. Navigating joint projects with dialogue. *Cognitive Science*, 27(2): 195–225, 2003.

[5] J. Bear, J. Dowding, and E. Shriberg. 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, 1992.

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

[7] Susan E. Brennan. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Lingusitics*, 2000.

[8] Susan E. Brennan and Eric A. Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, 8: 143–151, 1995.

[9] Janet E. Cahn and Susan E. Brennan. 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, 1999.

[10] H.H. Clark and Susan E. Brennan. Grounding in communication. In J. Levine L.B. Resnik and S.D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. 1991.

[11] H.H. Clark and E.F. Schaefer. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2: 19–41, 1987.

[12] H.H. Clark and E.F. Schaefer. Contributing to discourse. *Cognitive Science*, 13: 259–294, 1989.

[13] M. Core and L. Schubert. Implementing parser metarules that handle speech repairs and other disruptions. In *Proceedings of the 11th Annual International FLAIRS conference*, 1998.

[14] Peter A. Heeman and James Allen. Tagging speech repairs. In *Proceedings of the ARPA workshop on human language technology*, 1994.

[15] Emiel Krahmer, Marc Swerts, Mariet Theune, and Mieke Weegels. Error detection in spoken human-machine interaction. In *Proceedings of Eurospeech'99*, Budapest, Hungary, 1999.

[16] Emiel Krahmer, Marc Swerts, Mariet Theune, and Mieke Weegels. Problem spotting in human-machine interaction. In *Proceedings of Eurospeech'99*, Budapest, Hungary, 1999.

[17] Tim Paek and Eric Horvitz. 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*, 1999.

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

[19] Matthew Purver. A clarification request markup scheme for the BNC. Technical Report TR-02-02, Department of Computer Science, King's College London, February 2002.

[20] Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In R. Smith and J. van Kuppvelt, editors, *Current and New Directions in Discourse and Dialogue.* Kluwer Academic Publishers, 2002.

[21] David Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

Appendix 1: BNC files used in this study.

Filename	Number of Lines
KM8	1,222
KN3	875
KNC	1,310
KND	808
KNF	772
KNR	455
KNS	518
KNV	1,247
KNY	2,414
KP0	1,081
KP1	9,864
KP2	1,078
KP3	2,902
KP4	4,378
KP5	4,101
KP6	3,659
KP7	374
KP8	3,884
KP9	1,211
KPA	3,735
KPB	620
KPD	806
KPE	3,201
KPF	538
KPG	6,804
KPH	1,694
KPJ	575
KPK	982
KPL	868
KPM	1,399
KPN	599
KPP	1,320
KPR	1,981
KPT	1,350
KPU	2,955
KPV	7,965
KPW	1,023
KPX	1,126
KPY	1,080
KR0	2,659
KR1	728
KR2	1,655
KRF	1,130
KRG	1,700
KRH	5,168
KRL	5,450
KRM	3,095
KRP	1,904
KRT	6,640
KRY	532
KS1	903
KS7	1,482
KSN	2,421
KSR	1,676
KSS	5,147
KST	5,346
KSU	494
KSV	6,014
KSW	1,099
Total	**138,017**