Problems for Reliable Discourse Coding Systems

Sherri L. Condon and Claude G. Cech

Department of English and Department of Psychology
Center for Advanced Computer Studies
Université des Acadiens
University of Southwestern Louisiana
Lafayette, LA 70504
[slc6859,cech]@usl.edu

Abstract

Focusing on issues of intercoder reliability, this paper describes problems experienced in designing coding systems that classify language using discourse-relevant categories. First, given the absence of a consensus among language scholars, we examine options for selecting and structuring code categories, particularly those which have an impact on intercoder reliability. We observe that computer-assisted coding can maximize the options available to researchers for selecting and structuring code categories as well as minimize the problems of achieving and evaluating intercoder reliability. Second, focusing on the latter, we identify three alternative measures of intercoder reliability and present data from reliability tests using the strongest measure. These data show how structural properties of categories, such as their frequency or their status as a second pair part, can influence intercoder reliability. Despite the need to exercise caution in interpreting and reporting the results of reliability testing, researchers should view development of coding systems like development of theories: as a dynamic process in which reliability tests may provide yet further opportunities for theoretical validation.

Introduction

In the humanities, then in the social sciences, and more recently in the information sciences, researchers have sought to classify language into discourse-relevant categories. Consequently, most have grappled with the same fundamental questions that arise whenever linguistic forms are investigated in the context of human behavior: what is the relation between a linguistic form and the meaning(s) or function(s) with which it is associated? Through what procedure is this relation satisfied by a given form in a given context? Does such a relation or procedure even exist? It appears that any attempt to develop a discourse coding system is doubly burdened. Not only would the system presume solutions to problems that continue to challenge researchers in semantics and pragmatics, but also it must be one which can be used reliably by any individual who has received some reasonable amount of training.

These two basic problems of coding discourse units into discourse-relevant categories can be illustrated by comparing the coding of word units using syntactic categories. When classifying words into syntactic categories such as noun, verb and adjective, researchers can be confident they have selected categories which not only have a long history of significance in syntactic theory, but also have currency in most modern work. In contrast, no such tradition exists for the categories that researchers observe in extended discourse. While modern parts of speech still bear a remarkable resemblance to those proposed by the grammarians of ancient Greece (not to mention the Arab tradition and others), no one suggests that researchers devise discourse coding systems based on the detailed categories identified in classical rhetoric. Furthermore, when researchers in discourse survey current approaches to language in context, they encounter a myriad of discourse-relevant categories from the classification of genres in literary theory to the Freudian slip. The diversity reflected in Mann & Thompson's (1992) collection of 12 papers analyzing the same text represents only a fraction of the many approaches that have been adopted by ethnographers (Hymes 1962), sociologists (Goffman 1981), philosophers (Searle 1975), historians (Foucault 1972), and linguists (Halliday & Hassan 1976, Sinclair & Coulthard 1975, Tannen 1989, van Dijk 1981, and many others). Therefore, the first fundamental problem encountered in development of discourse coding systems is the selection of categories in the absence of clear consensus among language scholars.

The second problem, shared by both syntactic and discourse coding, is that no matter how well-chosen the categories are, the coding process requires more than tacit

knowledge of the language. As anyone who has taught grammar knows, though people effortlessly use nouns and verbs every day, they may not be able to identify nouns and verbs in a text. Similarly, even the most well-motivated discourse categories will not necessarily be easy to identify, which compounds the problem of developing a system that can be used reliably by multiple coders. Reducing code categories to a reasonably small number increases the likelihood that language will be encountered which fails to match any category, and sometimes it just isn't clear how a given piece of language should be classified. For example, theorists might disagree on a syntactic category for *far* and *way* in *far too many people* and *way too much beer*. But a coding system is capable of establishing a consistent, if arbitrary, convention for coding the forms by creating a new category, associating the forms with some pre-existing category, or dumping them into what our coders call a "garbage can" category, a category expressly for forms that do not fit into any other category. While developers of coding systems prefer to minimize arbitrary classifications, at least they have that option, which is not available to theorists. Though some portions of the coding system might not be well-motivated, the system at least guarantees--in principal--that all similar forms will be treated in the same fashion. Therefore, the value of a coding system requires ensuring that coders use the system reliably. At the same time, once reliable categories have been identified, theory-relevant questions (questions of validity) may be addressed.

Selecting Code Categories

Because the need to develop a reliable coding system is so important, the reliability problem cannot be treated as if it were separate from the problem of selecting code categories. However, the concern for reliability is only one of many issues encountered in selecting and structuring code categories. Often researchers are seeking to operationalize theoretical concepts and their choices are limited by these concerns. On the other hand, if the concepts are abstract enough, they can be associated with several different kinds of categories to bolster the validity of the operationalization. For example, Kiesler et al. (1985) evaluate uninhibited behavior using categories that they describe as follows: "impolite remarks," "swearing," "explicit statement of high-positive regard," "outbursts," and "superlatives" (94). Therefore, their measures of uninhibited behavior are based on properties of the discourse which are pragmatic (politeness and outbursts), semantic (statement of high-positive regard), lexical (swearing), and morphosyntactic (superlatives).

One would expect that categories defined by formal and structural properties would be coded more reliably than those which require semantic and pragmatic decisions, since the latter can be difficult to state precisely. For example, swearing and superlatives require only the determination of whether a word or phrase belongs to a fairly small set of items or whether it includes one of two structural relations to specific morphemes. In contrast, it is not so clear how one would explicitly define impoliteness, outbursts, and statements of high-positive regard. Of course, if a category is defined in a genuinely explicit manner, then it should be possible to eliminate human coding errors altogether with machine coding, and, in fact, some research compares discourse genres using syntactic and lexical information available in a tagged corpus (Biber 1988).

However, in practice, there are solid limitations to the coding that machines can do. For example, the category "discourse marker" in the coding system described in Condon & Cech (1992) has proven to be identifiable with high levels of intercoder reliability (90-100%, see below) in part because it is possible to provide a list of the forms (*ok, well, so, anyway, let's see,* etc.) that coders should identify as discourse markers. However, the decision also requires coders to discriminate between the discourse marker function of these forms and other functions served by the same forms, such as the use of *ok* to signal agreement and compliance, the adverbial function of *well,* and the use of *so* as a conjunction. The high levels of reliability are a result of the fact that these discriminations, which would be quite difficult to describe in an algorithm, are relatively easy for human coders to make.

Nevertheless, the possibility of computer-assisted coding promises to increase greatly researchers' options for designing and implementing coding systems. Lampert & Ervin-Tripp (1993) recommend using software that "prompts the user for individual fields and checks their entered responses against a set of allowable values for that field" (195). Computer-assisted coding makes it possible to design coding systems like Ervin-Tripp's with many categories and subcategories without concern for the limitations of coders' memories. It would allow the coding system to include categories defined extensionally by lists of items that are too long for human coders to remember and check efficiently. In fact, coding software could be designed to allow on-line checking of category definitions, to ensure that coders follow prescribed coding procedures, and to help prevent coders from overlooking infrequent categories. The possibility of machine assisted coding is an exciting one that researchers cannot afford to ignore.

While categories identified by formal and structural properties are likely to be easier to define precisely, categories based on semantic and pragmatic interpretations of the language can be quite reliable. For example, students are generally good at identifying instances of sense relations, such as synonymy and contradiction, on

exams. Kellerman et al. (1989) report very highs levels of intercoder reliability (96.7%) for a system which requires coders to isolate sequences of utterances "dealing with a single topic and having a single overarching content objective" (52). They used discourses produced when dyads who were unacquainted with each other were instructed to converse as they normally would when informally meeting someone for the first time, and coders were able to identify sequences which exchanged information about where they live, their work, their education, and so on. The judgment of whether speakers talked about a particular topic is evidently a clear one for coders, in part, perhaps, because specific lexical items such as *work* and *school* can play a role in identification of the sequences.

The system devised by Kellerman et al. avoids many problems by not requiring that a category be associated with each utterance or turn unit. However, for most quantitative methodologies, there is a need to establish both a coding unit and a set or sets of mutually exclusive and exhaustive categories into which the units are classified. It would be ideal if researchers could establish a standard unit that would afford some comparability across corpora along with a standard, all-purpose coding system, but in practice, these will be difficult to achieve. While the conversational turn is a natural one for spoken discourse and functional categories, it is not a natural one for written discourse and syntactic analyses. In fact, the coding units adopted by researchers vary almost as much as their code categories do. Condon & Cech (in press) define an utterance unit using a basic definition that applies to both oral and written discourse: the matrix clause together with all complements and adjuncts, including sentential ones. Thus clauses connected by coordinating conjunctions such as *and* and *but* were separated, but clauses connected by subordinating conjunctions such as *if* and *when* were not. Then additional conventions can be applied for units that are smaller than a clause, such as intonation and turn units in the oral interactions and the writer's choice of punctuation and spacing in the written, computer-mediated interactions.

The system has worked well for our purposes, but our failure to separate clauses connected by subordinating conjunctions would not be satisfactory for analyses like those of Mann & Thompson (1989), in which the relations expressed by those conjunctions must be identified separately. On the other hand, if Condon & Cech were to establish the smaller coding units required by Mann & Thompson, the primary effect would simply be to increase the number of utterances coded in our garbage can categories because our coding system does not discriminate the functions established by the additional units that would be coded. Therefore, it is certainly possible to establish a useful coding unit that could be identified in a broad range of discourses with no more difficulty than the identification of word units, which, of course, is not always simple. The problem is that the unit would likely be smaller than many researchers' interests require.

Of course, the same considerations hold for any attempt to standardize an all-purpose coding system as well. The real problem seems to be inherent in any approach based on function and the fact that at some point, discourse analysts must consider the functions of utterances in human behavior. The inherent problem for functional analyses is that there seems to be no limit to the functions that can be identified in language and, consequently, to the functions that can be associated with any given unit of language: Jane says hello to greet Harry, which functions to initiate a conversation with Harry, which functions in part to acquire information from Harry, which helps Jane determine whether he is interested in going out with her, which allows her to modify her strategies for obtaining a date with him, which satisfies her need or desire to interact socially with men, which functions as a preliminary to mating, which preserves the species, and so on. Therefore, not only the diversity of approaches to discourse but also the diversity of approaches to human behavior should preclude the possibility of developing an all-purpose utterance-unit coding system in the near future.

Instead, each researcher decides which functions will be examined based on his or her interests and theoretical alliances. With such an approach, what will be crucial for validating the code system is the extent to which the system is informative about the particular theory and the extent to which the theory is viewed as being informative about discourse. In our research, we are interested in how participants achieve understandings in interaction, but which of (infinitely?) many possible types of "understandings" should we investigate? In the methodology adopted for Condon (1986), Condon & Cech (in press) and related work using a similar coding system (Condon, Cooper, & Grotevant 1984; Cooper, Grotevant, & Condon 1983; Cooper, Grotevant, & Condon 1982), participants engage in simple decision-making tasks such as planning a weekend getaway, and we focus on the understandings that they must achieve in order to complete the task. They must be able to generate suggestions, evaluate those suggestions and determine from the evaluations whether the suggestion acquires the status of a decision. By focusing on these and similar functions, we overlook many others present in the discourse, but avoid much of the subjectivity that can be associated with functional analyses.

There are many additional considerations that researchers must consider when developing coding systems, and the reader should consult Lampert & Ervin-Tripp (1993) for an excellent discussion. A final concern is elaborated here because of its relevance to the problem of establishing and evaluating reliability. Since a given unit of

language may serve many functions, odds are that even though a particular researcher is only interested in a small subset of these functions, some units will serve more than one of the relevant functions. For example, (1a) both orients a suggestion and requests information, while (1b) both complies with a request for information and formulates a suggestion.

(1) a. what do you want to do in the morning
b. sleep

The problem can be resolved in part by creating many sets of mutually exclusive categories made exhaustive by liberal use of garbage can categories. For example, Condon & Cech (in press) establish three sets of categories: MOVE functions (suggests, requests action, requests information), RESPONSE functions (agrees, disagrees, complies with request), and OTHER functions (discourse marker, orientation, metalanguage), each with its own garbage can, "no clear MOVE/RESPONSE/OTHER function."

Still, some utterances can be associated with more than one MOVE, RESPONSE, or OTHER category. For example, one participant in saying "Put Hawaii" both formulated a suggestion by introducing the idea of going to Hawaii for the first time and formulated a request for action by requesting that Hawaii be written on the answer sheet (which was given to participants to ensure that they made a minimal number of decisions). To preserve the mutual exclusivity within the MOVE, RESPONSE and OTHER sets of categories, the categories within each set were arranged hierarchically and coders were instructed to code the highest category on the hierarchy. Thus, suggestions were higher than requests for action, which in turn were higher than requests for information. The system proved to have the additional value of helping coders identify functions formulated in indirect language (Searle 1975). For example, it is well known that requests for action are often formulated in a way that would also allow them to be treated as requests for information (e.g. *Do you know what time it is?).* The "code high" convention recognizes this sort of possibility, encourages coders to identify those utterances with multiple functions and provides a simple procedure to determine a code category for them.

Yet in spite of our best efforts, we still encountered language that was difficult to code in our system. Consider the exchange in (2), which was produced by a family planning an imaginary 2-week vacation. The mother had asked *Do you want to start in England?* and while the father was replying *England or France or Spain,* the mother continued, saying that they could start in Spain. The father continued with (2a) while two children each exclaimed *Spain,* so there was a considerable amount of overlapping speech that ended with the mother's reply in (2b).

(2) a. but uh it depends on where you want to go
b. well ok England Spain Austria

The problem is whether (2a) should be coded as a statement about the decision process or as an indirect request for the mother to identify the places she wants to go. From the father's statement that deciding where to go is logically prior to deciding where to start, the mother could have inferred a plan to identify the places she wants to go as an appropriate response to the statement. On the other hand, she might have interpreted the statement as an indirect request for her to decide where to go in the same way that *It needs salt* can be interpreted as a request to provide salt.

Clearly, an account of what has happened in (2) would require answers to fundamental questions about language and human behavior. *Researchers should not overlook the similarity between a coding system and any theory or model that seeks to associate linguistic form with contextualized meaning.* Neither coders nor theorists have access to the mental processes that accompanied the exchange in (2), but the fact is neither do speakers have access to the mental processes of hearers. This evidently does not prevent speakers from reasoning about the mental states of hearers to formulate appropriate continuations of the discourse, which is exactly what many researchers would like to understand. Therefore, we think of exchanges like (2) as the kind of recalcitrant data that both challenges our theories and points out the directions we need to explore. While they have the effect of decreasing the reliability of our coding system, the problem need not lie in the coding system itself.

Evaluating Intercoder Reliability

Some problems of establishing intercoder reliability will be minimized by computer-assisted coding: the limitations on human memory, the failure of coders to follow prescribed procedures, perhaps even some of the response bias toward frequent categories can be eliminated in creative ways. Other problems, such as the indeterminacy described above with respect to the data in (2) and the general tendency for categories to drift with context (cf. Barsalou 1987; Cech, Shoben, & Love 1990), are likely to prove more intransigent. Although it is inefficient, researchers who seek to maximize their confidence that the language is coded as they intend will have each discourse coded by more than one coder with differences resolved by those coders and/or another coder. However, they are still faced with the problem of evaluating and reporting the intercoder reliability they achieve.

First, there are several ways of measuring intercoder reliability, including a weak measure that considers only the total number of items coded in each category, not whether coders associated those categories with the same units. Similarly, unit-by-unit measures of intercoder agreement are still weaker than establishing a separate "standard" coding of the discourse and measuring each

coder's agreement with that standard. The strongest measure is important if the system includes difficult or infrequent categories that coders might overlook, and it also helps to avoid drift. Furthermore, when reliability is evaluated in this manner, it is possible to observe how individual coders and individual discourses influence the measures. The results presented below were all obtained using the strongest measure.

Second, it is difficult even to select material for reliability testing. Researchers who require more than one coder to code each transcript have the option of tracking intercoder agreement for each unit of each discourse by having coders keep a record of every disagreement encountered during coding. Of course, this procedure would not provide the strongest measure, and unless the process were implemented in a computer-assisted coding system, it would be extremely inefficient both to record every disagreement and to calculate every measure. Consequently, the problem of selecting material remains and increases for researchers who want to develop standard samples for the strongest measure of intercoder reliability.

Because utterance-unit discourse coding must be done in context, samples must be stretches of continuous discourse. They cannot be selected randomly because they must satisfy so many criteria: they must be reasonably small, yet include a representative sample of all code categories. Ideally, samples should come from beginnings, middles, and ends of discourses and from many different types of discourses with many different interlocutors. In practice, it can be difficult to find reasonably small stretches of discourse that include even one exemplar of every code category if the code includes categories that occur infrequently. The alternative of slicing out smaller portions of text containing infrequent categories runs the risk of obtaining low levels of agreement because small portions of the discourse provide insufficient context for accurate coding.

Third, infrequently occurring categories present other problems as well, beginning with the paradoxical fact that the researcher wants samples to be representative of the data but, since the categories occur infrequently, truly representative samples will have few exemplars of them. Thus as samples are selected to maximize opportunities for evaluating infrequently occurring categories, they become less representative of the data and, in the worst case, may include higher levels of unusual language that is difficult to code. Furthermore, infrequently occurring categories are more vulnerable to limitations on human memory and the possibility that coders will develop a response bias in favor of frequently occurring categories. But the problem that makes it most difficult to be confident about evaluations of infrequent categories is the fact that in practice one never obtains enough exemplars of the categories compared to those which occur more frequently. Statistically, coder identification rates may be regarded as estimates of the 'true' identification rates. The more opportunities for identifying a given code category, of course, the better the estimate. Therefore, researchers should not be surprised to experience difficulty obtaining high intercoder reliability for infrequently occurring categories.

To obtain an example of the effect of category frequency on coder agreement with a standard sample, we used records of reliability tests conducted for the face-to-face interactions analyzed in Condon & Cech (in press). Each test consisted of 60-80 utterances, which, like all interactions that coders received, had been transcribed and separated into utterance units. A coding of the transcripts by the first author was employed as the standard. Reasoning that the effects will be most evident before coders become fully competent with the system, we selected 5 tests consisting of the following cases: 4 were tests administered early in training to 3 excellent coders and 1 poor coder who never satisfied our criteria for coding. The 5th was a later test from the same coder.

Categories were identified as frequent if they occurred at least at half the rate of the most frequently occurring category. Thus, because suggestions and elaborations (a catch-all MOVE category) occurred most frequently with 64 and 65 respectively, we identified frequent categories as those for which there were more than 32 exemplars. They consisted of 5 categories, including requests for information, agreements, and discourse markers along with the two mentioned above. The infrequent categories consisted of 9 other categories, including two MOVE categories (requests for action, requests for validation) and all of the remaining RESPONSE and OTHER categories (disagreements, compliance with requests, acknowledgements, metalanguage, orientation, exchanges of personal information, and jokes or exaggeration). Summing across the frequent and infrequent categories, we find 77% agreement with the standard for the frequent categories and only 58% for the infrequent categories. Since there were so many more categories in the infrequent group, we also summed across the 5 most infrequently occurring categories in the sample and obtained 59% agreement for that group. Finally, since the results might be attributable to inclusion of a poor coder, we also recalculated the measures excluding the results from that coder. The revised measures showed 76% agreement with the standard for frequent categories and 61% agreement with the standard for the infrequent categories. For the 5 least frequent categories, there was only 50% agreement with the standard when the poor coder was excluded. Consequently, in initial reliability testing, the problems of infrequently occurring categories can make it appear that they are being identified at low levels.

However, infrequently occurring categories can be extremely important for researchers' theories and methodologies, as the discussion of the exchange in (2) above

suggests. For example, we are interested in how interactants manage decision-making discourses, but only one of the 4 categories most relevant to discourse management in our data is one of the 5 frequently occurring categories identified above. Furthermore, infrequently occurring categories can be quite salient to coders, as we experienced with the two categories designed to access affective domains of the discourse (exchanges of personal information and joking or exaggerating). Finally, as good coders become more experienced, these differences can be reduced. We calculated another measure for frequent and infrequent categories using a sample of 4 later tests from the 3 excellent coders who coded the face-to-face interactions in our data. A problem immediately arose when the frequencies of the categories in the two samples did not match. The 5 categories that were most frequent in the first sample continued to be among the 6 most frequent categories in the second sample, but two of the 5 (agreements and discourse markers) were less frequent than a sixth category, compliance with requests, and all three of these occurred with slightly less than half the frequency of the most frequently occurring categories (which were again suggestions at 48 and elaborations at 54). Consequently, we compared the 6 most frequent categories to the remaining categories and obtained 87% agreement with the standard for both groups. This suggests that the difficulties represented by infrequently occurring code categories can be minimized, but a second measure illustrates how pervasive they are: when we compared the 5 least frequent categories in the second sample, we obtained only 71% agreement with the standard.

Reliability and Theory-Based Validation

We have suggested that coding systems are valued most for their reliability, and need not demonstrate validity. Nevertheless, in some instances, tests of reliability may be informative about the underlying theory, so that reliability assessment may simultaneously perform a validating function. One such instance in our own work concerns the fact that certain code categories should be related, such as our MOVE and RESPONSE categories. The second pair parts of adjacency pairs (Goffman 1981; Schegloff & Sacks 1973) exist only as responses to first pair parts. If coders fail to identify a first pair part, they should not be able to identify the second pair part, if the *theorized* dependency of the latter on the former is in fact valid. Thus, as identifications of MOVE categories become more reliable, so should identifications of RESPONSE categories. Summing across 5 MOVE categories and 4 RESPONSE categories, we found 74% agreement with the standard for MOVE categories in our early sample and 85% in our later sample. Consistent with what our theory would lead us to expect, RESPONSE categories exhibited a similar pattern: 58% for the early sample, but 87% for the later sample. In short, when MOVE categories are identified with a high degree of reliability, so are their corresponding RESPONSE categories; but as MOVE functions become harder to identify, identification of the RESPONSE functions suffers correspondingly more.

Of course, MOVE categories occur more frequently than RESPONSE categories, so these results are tentative, but they suggest that data from reliability testing can function to provide additional validation of the theories which inform the coding system. If the claim that RESPONSE categories are dependent on MOVE categories is wrong, then RESPONSE categories might be independent of MOVE categories and so be identified independently as well. We can test the model by asking whether the probability of correctly identifying a RESPONSE category systematically varies when conditionalized on correct identification of the MOVE category. Therefore, a model of discourse categories may also make predictions concerning the behavior (i.e., the reliability) of coders. A rather unusual consequence of this type of reasoning is that it may not be a good idea to train coders to a criterion of 100% reliability. We have already described the problems inherent in low-frequency categories: to ensure that coders have reached criterion for these, either overtraining will actually have occurred on the high-frequency categories, or the infrequent categories will have been presented during training at frequencies that are disproportionate to their frequencies in the actual discourse. Now we can see that 100% reliability also precludes the possibility of testing predictions about variation in coder agreement. Consequently, instead of viewing the achievement of reliability as a Promethean quest for impossible perfection, researchers can look for opportunities to employ reliability assessments to test theoretical claims.

Clearly, researchers must use good judgment in evaluating and reporting the intercoder reliability of discourse coding systems. Reliability may decrease in some cases for reasons that are theoretically informative (such as with RESPONSE categories), and in other cases for reasons having to do with adequacy of statistical samples (such as with infrequent categories). Researchers should not be disappointed when they experience these sorts of problems in evaluating intercoder reliability given the multiple factors that make those evaluations quite complicated and tentative. Rather, such 'obstacles' ought to be viewed as occasions for further theoretical elaboration and evaluation, including (as demonstrated above) theorized dependencies between code categories.

Summary

Some of the best advice that Lampert & Ervin-Tripp (1993) offer is their suggestion that "the construction of a coding system should be seen as a dynamic and ongoing

process that benefits at the start from original theories and research goals, yet has enough flexibility so that later revisions can be made to capture unexpected and interesting distinctions" (182). We have observed that developing a discourse coding system can be much like developing a theory of discourse with the added problem of achieving high levels of intercoder reliability. Many factors create problems for evaluating intercoder reliability, including such structural factors as the frequency of occurrence of the category and its status as a first or second pair part. We have demonstrated that the effects of these factors appear to weaken as coders become more experienced. Furthermore, we have observed the potential for using reliability assessment to test theoretical claims. Finally, we conclude that computer-assisted coding provides some exciting options for designing coding systems and increasing the reliability with which they are applied.

Acknowledgments

This project was supported by a Faculty Research Award from the University of Southwestern Louisiana. For their help with data analysis, we gratefully acknowledge Cathy Landry, Joyce Lane, Tom Petitjean, and Traci Smrcka.

References

Biber, D. 1988. *Variation Across Speech and Writing.* Cambridge University Press.

Barsalou, L. 1987. The instability of graded structure: Implications for the nature of concepts. In Neisser, U., ed., *Concepts and Conceptual Development.* Cambridge University Press.

Cech, C.; Shoben, E.; and Love, M. 1990. Multiple congruity effects in judgments of magnitude. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16: 1142-1152.

Condon, S. 1986. The discourse functions of OK. *Semiotica* 60: 73-101.

Condon, S. and Cech, C. 1992. Manual for Coding Decision-Making Interactions. Unpublished manuscript, Lafayette, La.: Université des Acadiens (University of Southwestern Louisiana).

Condon, S. and Cech, C. 1995. Functional comparison offace-to-face and computer-mediated decision-making interactions. In Herring, S., ed., *Computer-Mediated Communication.* Philadelphia: John Benjamins (in press).

Condon, S.; Cooper, C.; and Grotevant, H. 1984. Manual for the analysis of family discourse. *Psychological Documents* 14(1): Document no. 2616.

Cooper, C.; Grotevant, H.; and Condon, S. 1982. Methodological challenges of selectivity in family interaction: addressing temporal patterns of individuation." *Journal of Marriage and the Family* 44: 749-754.

Cooper, C.; Grotevant, H.; and Condon, S. 1983. Individuality and connectedness in the family as a context for adolescent identity formation and role-taking skill. In Grotevant, H. and Cooper, C., eds., *Adolescent Development in the Family.* San Francisco: Jossey-Bass.

Foucault, M. 1972. *The Archaeology of Knowledge.* New York: Pantheon Books.

Goffman, I. 1981. *Forms of Talk.* University of Pennsylvania Press.

Halliday, M. and Hassan, R. 1976. *Cohesion in English.* London: Longman.

Hymes, D. 1962. The ethnography of speaking. In Gladwin, T. and Sturtevant, W., eds., *Anthropology and Human Behavior.* Anthropological Society of Washington.

Kellerman, K.; Broetzmann, S.; Lim, T.; and Kitao, K. 1989. The conversation MOP: Scenes in the stream of discourse. *Discourse Processes,* 12: 27-61.

Kiesler, S.; Zubrow, D.; Moses, A.; and V. Geller, V. 1985. Affect in computer-mediated communication: an experiment in synchronous terminal-to-terminal discussion. *Human-Computer Interaction* 1: 77-104.

Lampert, C. and Ervin-Tripp, S. 1993. Structured coding for the study of language and social interaction. In Lampert, C. and Ervin-Tripp, S., eds., *Talking Data: Transcription and Coding in Spoken Discourse.* Hillsdale, NJ: Lawrence-Erlbaum.

Mann, W. and Thompson, S. 1989. Rhetorical structure theory: A theory of text organization. In Polanyi, L., ed., *The Structure of Discourse.* Norwood, NJ: Ablex.

Mann, W. and Thompson, S. 1992. *Approaches to Discourse.* Philadelphia: John Benjamins.

Schegloff, E. and Sacks, H. 1973. Opening up closings. *Semiotica* 8: 289-327.

Searle, J. 1975. Indirect Speech Acts." In Cole, P. and Morgan, J., eds.,*Syntax and Semantics.* Vol. 3: *Speech Acts.* New York: Academic Press.

Sinclair, J. and Coulthard, R. 1975. *Towards an Analysis of Discourse.* Oxford University Press

Tannen, D. 1989. *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse..* Cambridge University Press.

van Dijk, T. 1981. *Studies in the Pragmatics of Discourse.* The Hague: Mouton.