A METHOD FOR MEASURING CONVERSATIONAL COHERENCE

Alex Black*

Department of Psychology
University of Victoria

1. INTRODUCTION

The phenomenon of coherence is of much interest to linguists, psychologists, text analysts, and communication scholars. Indeed there are many theories about the cognitive and social factors involved in the production and comprehension of a coherent conversation or text (McLaughlin 1984). However, as Hobbs (1979) noted, these theories often amount to nothing more than taxonomies which have no empirical basis. In part, the paucity of experimental validation stems from a lack of valid and reliable measure of coherence. In this paper, current measures of coherence are reviewed, and the reliability and validity of a new measure of coherence is assessed.

2. PREVIOUS MEASURES OF CONVERSATIONAL COHERENCE

Generally, four methods have been used in the assessment of conversational coherence: (1) judgement by fiat; (2) restructuring of conversations (Ellis, Hamilton and Aho 1983); (3) judgement of appropriateness (Phanlap and Tracy 1980); and (4) scaling of equivocation (Bavelas and Smith 1982).

Most scholars use the "judgement by fiat" technique to assess coherence (Hobbs 1979). Typically, these scholars analyze a conversation in terms of a taxonomy of coherence relations that they propose. Usually, the rules for making decisions about the types of coherence are included in the analysis, and the types of coherence relations are based on the semantic content of the statements (e.g. Riechman 1978, Shank 1977). None of these scholars, however, have attempted to establish the reliability of their methods of assessment. Therefore, while this technique may be a valid measure of coherence (as defined by the particular theory of coherence), its reliability has yet to be established. In addition, the "judgement by fiat" method of measuring conversational coherence has an objectivity problem; that is, the people who classify conversations are the very scholars who are presenting the theory of coherence.

In testing the validity of Reichman's (1978) model of conversational coherence, Phanlap and Tracy (1980) asked subjects to judge whether the second statement of a pair was an appropriate response to the first statement. If we consider coherence to be the semantic relation between a pair of sentences (McLaughlin 1984), it is important to note

---

* The author would like to thank Janet Bavelas, Charles Lemery, and Jennifer Mullett. Without their tolerance, advice and continued enthusiasm this measurement method would not have been developed. The author would also like to thank Nicole Chovil and Dianna MacGibbon for their thoughtful and appropriate editorial comments.

that this procedure does not operationalize the most common view of coherence. Subjects were asked to judge the appropriateness of a response and not the semantic relations between two statements. Because there is no empirical evidence linking coherence with the judgements of appropriateness, it is difficult to judge whether the measurement technique actually assesses coherence. The coherence measure used by Ellis, Hamilton and Aho (1983) has a similar limitation. Ellis *et al.* (1983) transcribed a conversation, printed the statements on cards, randomly ordered the cards, and asked subjects to sort the cards in the order the conversation occurred. Since the subjects were not instructed to base their judgements on the coherence between the statements, it is difficult to judge what criteria the subjects used in sorting the cards. Therefore, while the procedure of Ellis *et al.* (1983) is a valid measure of how subjects think conversations should be structured, there is no evidence that the subjects' sorting of the conversation reflects the coherence between the statements.

Finally, Bavelas and Smith (1982) have developed a reliable and valid measure of equivocation that, in part, measures the degree of conversational coherence occurring between questions and answers. As described in Bavelas and Smith (1982) students are trained over a series of scaling sessions to rate the extent to which a speaker's utterance: (1) is clear; (2) addresses the other person in the situation; (3) contains the speaker's opinion; and (4) is a direct answer to the question. The final dimension (extent to which a message is a direct answer to the question) can be seen as a valid measure of the degree of coherence occurring between questions and answers. Moreover, as reported in subsequent articles (Bavelas 1985, Bavelas and Chovil 1985) the procedure has proven to be a reliable measure for distinguishing between coherent and incoherent messages. Unfortunately, the equivocation measurement technique is unsuitable for the general measurement of coherence, because the last dimension can only assess coherence between questions and answers.

In summary, current measures of conversational coherence are limited. Phanlap and Tracy's (1980) and Ellis, Hamilton and Aho's (1983) measures of coherence are questionable, because properties other than the semantic relations between statements are assessed. The judgement by fiat technique of coherence has no demonstrated reliability, and Bavelas and Smith's (1982) equivocation measure can only assess the coherence between questions and answers.

3. A METHOD FOR MEASURING THE DEGREE OF COHERENCE

3.1 Overview

Naive judges rated the degree of semantic similarity between a pair of statements on a "coherence" continuum using a magnitude estimation procedure similar to that of Bavelas and Smith (1982). This procedure avoids the limitations of other measures of coherence. First, the lay judges are not aware of the theory in question. Second, the degree of semantic similarity of two statements is assessed; therefore, the measurement technique operationalizes a more common definition of coherence. Finally, the measurement procedure assesses the coherence between all types of statements.

3.2 Stimuli and the Coherence Continuum

The judges rated the semantic similarity between 55 pairs of videotaped statements. The statements were obtained from a short excerpt of an interview between Barbara Frum and John Turner that occurred during the Liberal Party of Canada leadership campaign (see Table 1). The pool of 55 statements was formed by combining the eleven

statements into all possible nonredundant pairs. The pairs of statements were edited into a stimulus tape where each pair of statements was preceded by an identifying number. (Copies of the edited videotape are available from the author.)

Judges rated the stimuli on a piece of masking tape placed across the length of a desk (120 cm). Four labels were placed underneath the masking tape in order to convey the continuous nature of the coherence continuum: at the extreme left, the label read, "This area is for messages that are very connected. Messages like this are about the same things."; at the extreme right the label read "This area is for messages that are disconnected. Messages like this do not have a general topic in common." At 40 and 80 cm were labels that read "This area is for messages that are quite connected. Messages like this are related in that they are talking about very similar things." and "This area is for messages that are quite disconnected. Messages like this may share a general topic, but are only indirectly related." Thus, judges rated the degree of semantic similarity between statements on a physical scale of magnitude. Statements with high semantic similarity were rated at the far left-hand side of the continuum, while statements with no semantic similarity were rated at the far right-hand side of the continuum.

3.3 Judges

Four University of Victoria undergraduate students participated in the study. Each judge was paid five dollars.

3.4 Instructions and Procedure

Judges met with the experimenter separately. At the outset the judges were told that the investigator was interested in the "connectedness of utterances" and that their job was to rate the "relatedness between some messages". The experimenter then described the continuum and played examples of two very related messages, two related messages, two messages that were indirectly related, and two messages that were not at all related (see Table 1 for transcription of the examples). After the experimenter presented the videotaped examples and stressed that the ratings should reflect literal interpretations of the messages, he described the messages that would be rated and instructed the subjects to make a pencil mark on the tape for each rating.

The experimenter then played the stimulus videotape for each judge who rated the pair of statements, identified it by number, and then proceeded to the next pair of statements. If the judge had trouble hearing the statements or had difficulty making a judgement, the experimenter replayed the pair of statements. At all times the judges were encouraged to talk about their ratings, so that the experimenter would know what criteria the judgements were based on.

3.5 Results

In order to estimate the reliability of the judges' ratings of the 55 pairs of statements, an intraclass correlation (Ebel 1951) was calculated. The intraclass reliability of the students ratings was high ($r = 0.88$). Parenthetically, the judges did not find the scaling task to be difficult, and made very similar comments about the coherence of the pairs of messages.

One might dismiss the high reliability, because of the small N (4) used in the study. Actually, the use of a small N provides a more severe test of reliability, because one idiosyncratic judgement results in proportionally greater error variance within a small N set of judges than the same idiosyncratic judgement within a large N set of judges. Thus,

Table 1. Examples and Stimulus Statements.

Examples

1 (very connected). a) The agency did not tell us the extent of their involvement. b) They did not tell us the extent of the President's involvement.

2 (very unconnected). a) The agency did not tell us the extent of their involvement. b) I have not endorsed Senator Hart for several reasons -- the most important of which is I'm not sure of where Gary Hart is going either.

3 (quite connected). a) The agency did not tell us the extent of their involvement. b) What we were in fact briefed about, about two months after the fact, was that in a compodium of a number of things that were occurring in at number 17 was by the way there are mines there.

4 (quite disconnected). a) The agency did not tell us the extent of their involvement. b) Isn't it a fact that you're opposed to the contras being financed by U.S. funds in order to overthrow the Sandinista regime?

Stimulus Statements

1. Are you embarrassed by the revelations that reflect ill of your financial competence?
2. I've been in the .. I've been in the free market.
3. And you win some Barbara and you lose some.
4. If I hadn't won some I wouldn't have been able to stand for the office that I am standing for.
5. You are asking for confidence to manage the whole economy, not just one small business.
6. I didn't ahh manage the business' personally.
7. I was not the chief executive officer.
8. I chaired a board.
9. and ahhh on the small business side ahh small business has been savaged throughout 1981, 1982 1983.
10. And we were on the venture capital field.
11. Under Ontario Legislation that that subsidized it because it was last resort risk financing.

the use of more judges will only increase the reliability of a measure (Nunnally 1967).

4. EXAMINATION OF THE RATINGS

The final step of the measurement procedure was to average the judges' ratings for each pair of statements. This average score reflects the degree of coherence between two statements: a low score indicates a high degree of coherence between two statements, while a high value indicates incoherence between two statements.

Examination of the scores reveals that the judges' ratings are both sensitive and subtle. For example, the judges rated the semantic similarity between message three and four as high, while statements two and three were judged as only tenuously related. Moreover, the scores that resulted from the students ratings reflect more subtle differences in coherence. For example, statements six and seven specifically relate to John Turner's activities within a single organization, while statement eight concerns what activities John Turner did *not* perform within the same business. Here judges rated the coherence between statements six and seven as slightly greater than the coherence relation between statements seven and eight.

5. DISCUSSION AND USE OF THE COHERENCE SCORES

The goal of this research was to develop a reliable measure of conversational coherence. The results described here are consistent with this goal. Lay judges with little training rated the semantic similarity between messages with a high degree of reliability. Still the obvious question remains, "Is the measure valid?" Because the best evidence of construct validity results from use of the measure (Cronbach and Meehl, 1955), the author's current research is relevant.

The measurement technique has been used to test a homeostatic conception of conversational coherence. Briefly, a homeostatic conception of coherence posits that statements are organized so that the global coherence of a conversation is maximized, and that global incoherence is minimized. Such a conception of coherence does not require statements to be organized at a local level in a Grician (1975) fashion (that is, "be relevant"). On the contrary, a homeostatic conception of coherence predicts that statements that follow one another will not always be relevant to one another, and that this lack of relevance is inextricably involved in maintaining the coherence of a conversation as a whole. Thus, a homeostatic conception of coherence would predict that the order in which the statements within a conversation occurred would be one of the most coherent possible orderings of the statements.

In order to test the homeostatic conception of coherence a computer program which simulated the different possible orderings of a conversation was written. Briefly, the program determines the global coherence of each possible ordering of the statements within a conversation, compares the total coherence of each possible ordering of the statements to the total coherence of the ordering of the statements that occurred, and then records the number of possible orderings that are more coherent than the ordering of the statements that occurred. When the coherence ratings of the Frum-Turner interview were entered into the program, the results strongly confirmed the homeostatic conception of coherence. Of the 3,628,800 possible combinations of the 11 statements, 18,303 (.5%) were more coherent than the order of the eleven statements that occurred.

Therefore, 99.5% of the possible orderings of the statements in the Frum-Turner interview were less coherent than the order of the statements that occurred.

To return to the question of validity, the research concerning the homeostatic conception of coherence provides partial confirmation that the measurement technique is valid. Indeed, because the hypothesis makes an extreme prediction about the coherence of the total conversation and requires that the measure have little measurement error, the results of the simulation program provide strong evidence for the validity of the measure.

Obviously, the validity of a measure is determined over a series of investigations; hopefully, readers of this paper will employ this measurement technique to examine other aspects of conversational coherence. For example, if the coherence between all the statements within a conversation is measured it is possible to depict a conversation topographically. With such a "map" of conversation it would be possible to depict the clusterings of statements about a single topic or note where a conversational change occurred. Or, the coherence within a conversation could be depicted graphically. From these graphs it is possible to examine both equivocation and repair sequences (Vuchinich 1977), since one speech act involves a departure from coherence and the other involves a return to coherence. Other possible modifications would involve the stimuli presented to judges: the measurement technique could be applied to written or tape recorded messages, or the unit of analysis could be changed to a pair of speaker turns or a pair of paragraphs.

However, the measurement technique presented here is not without limitation. The scalings reflect a literal interpretation of the statements; that is, the judges were instructed not to infer what the speaker of the statement "meant". Thus, the ratings in no way reflect the speakers' perceptions of the conversational coherence. Such a limitation is important when conversants resort to the use of the conversational implicature. Here a listener would consider a response to be coherent; yet the present procedure would consider the response to be incoherent. Since the goal of this research was to develop a reliable measure of the coherence between statements, limitation of the judges' inferences was unavoidable. However, to researchers who want to assess what the phenomenological aspects of coherence are, the measure of coherence presented in this paper is not of use, and may not be adaptable for their purposes.

REFERENCES

Bavelas, J.B. (1985). A Situational Theory of Disqualification. Using Language to "Leave the Field". Pp. 189-211 in *Language and Social Situations.* J. Forgas, ed. New York: Springer-Verlag.

Bavelas, J.B. and B.J. Smith. (1982). A Method for Scaling Verbal Disqualification. *Human Communication Research* 8: 214-227.

Cronbach, L.J. and P.E. Meehl. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin* 52: 281-302.

Ebel, R.L. (1951). Estimation of Reliability Ratings. *Psychometrika* 16: 407-424.

Ellis, D.G., M. Hamilton and L. Aho. (1983). Some Issues in Conversational Coherence. *Human Communication Research* 9: 267-282.

Grice, H.P. (1975). Logic and Conversation. In *Syntax and Semantics* 3: 41-58. P. Cole and J.L. Morgan, eds.

Hobbs, J.R. (1979). Coherence and Coreference. *Cognitive Science* 3: 69-90.

McLaughlin, M.L. (1984). *Conversation: How Talk is Organized.* Beverly Hills, CA: Sage Publications.

Nunnally, J.L. (1967). *Psychometric Theory.* New York: McGraw-Hill.

Phanlap, S. and K. Tracy. (1980). Not to Change the Topic But...: A Cognitive Approach to the Management of Conversation. In *Communication Yearbook* 4: 237-259. D. Nimmo, ed.

Reichman, R. (1978). Conversational Coherence. *Cognitive Science* 2: 283-327.

Schank, R.C. (1977). Rules and Tipics in Conversation. *Cognitive Science* 1: 421-441.

Vuchinich, S. (1977). The Elements of Cohesion between Turns in Ordinary Conversation. *Semiotica* 20: 229-257.