Application of the Correct Information Unit Analysis to the Naturally Occurring Conversation of a Person With Aphasia

Mary L. Oelschlaeger*
John C. Thorne**
University of New Mexico
Albuquerque

The Correct Information Unit (CIU) analysis for measuring the communicative informativeness and efficiency of connected speech (Nicholas & Brookshire, 1993) was applied to the naturally occurring conversation of a person with moderate aphasia. Results indicated that, in this instance, reliable CIU measures could not be obtained. Intrarater reliability for CIU and %CIU was low, reaching only 72%, and interrater reliability was never greater than 63%. However, reliability of word counts was good. Post hoc analysis of rater disagreements in application of the CIU analysis revealed that the majority (72%) resulted from insufficiencies in the scoring rules that were originally designed to measure single speaker connected discourse. Two descriptive categories of disagreements were identified: interpretations of informativeness and absence of rules. The remaining 28% of disagreements were attributable to human error in the application of scoring rules. Comparison of findings with previous research and implications for future research are discussed.

KEY WORDS: aphasia, conversation, measurement, correct unit analysis

An ongoing challenge in clinical aphasiology has been the description of the language ability of individuals with aphasia. This challenge has been approached primarily through the use of standardized testing instruments (Goodglass & Kaplan, 1972; Kertesz, 1982; Porch, 1981, for example). However, the observation that everyday language ability may exceed that predicted by formal test scores has led researchers to develop other tools that better capture the communicative success of persons with aphasia (e.g., Blomert, Koster, Van Mier, & Kean, 1987; Holland, 1980; Lomas et al., 1989).

One recent approach toward this end is reported by Nicholas and Brookshire (1993). In their study, these authors describe the Correct Information Unit (CIU) analysis, a rule-based scoring system for measuring two aspects thought to be important to communicative success: the informativeness (the degree to which the speech of the individual imparts the intended message) and efficiency (the rate at which the message is produced) of language use. Using specific elicitation stimuli (6 picture description, 4 requests for personal and procedural

*Currently affiliated with Northern Arizona University, Flagstaff

**Currently affiliated with Albuquerque Public Schools, Albuquerque, NM

information), a "connected speech" sample is obtained. The CIU analysis of connected speech involves measuring the rate at which the speaker produced speech and combining it with a derived measure: the %CIU. The %CIU measure combines the total word count in the connected speech sample with the number of words in that count which meet the specific criteria necessary to be called a correct information unit. A correct information unit is defined as "words that are intelligible in context, accurate in relation to the picture(s) or topic, and relevant to and informative about the content of the picture(s) or the topic" (Nicholas & Brookshire, 1993, p. 348).

As part of their description of the CIU analysis, Nicholas and Brookshire offer psychometric information related to the reliability and stability of scoring. Specifically, intrarater and interrater reliability exceeded 90% (Nicholas & Brookshire, 1993), and scores were noted to be reasonably stable over time (Brookshire & Nicholas, 1994). These results led the authors to the conclusion that the CIU analysis "may yield stable baseline performance against which changes in connected speech with treatment or manipulation of experimental variables can be measured" (Nicholas & Brookshire, 1993, p. 344). The authors go on to point out, however, that the stability of an individual's performance should not be assumed and that the application of the CIU as a measure of change with treatment effects would require determination of stability over repeated assessments for any individual.

Since this initial report, a number of research studies (Brookshire & Nicholas, 1994, 1995; Doyle, Goda, & Spencer, 1995; Doyle, Tsironas, Goda, & Kalinyak, 1996; Nicholas & Brookshire,1995) using the CIU analysis have been conducted. Most relevant to this discussion is one reported by Doyle, Goda, and Spencer (1995). These authors attempted to expand the application of CIU analysis to discourse types other than those in Nicholas and Brookshire's (1993) original study. They added conversational elicitation tasks to the original 10 stimuli, noting that "measuring communicative informativeness and efficiency under conversational discourse conditions is perhaps the most ecologically valid means of determining the interpersonal verbal communication abilities of adults with aphasia" (p. 130).

Doyle et al.'s study compared the performance of 20 persons with aphasia on Nicholas and Brookshire's discourse tasks with their performance on two conversational discourse tasks. Conversational samples were collected in "simulated natural environments" (Doyle et al., 1995, p. 131) and were of two different types. The first sample was identified as "topic-open." Subjects and their conversational partners were instructed to "discuss anything they chose." The second sample was "topic-constrained"; subjects watched a 4.5-minute segment of a network news program and were asked to discuss its contents. Both sampling procedures were 7 minutes in length. The %CIU of conversational samples was then compared with that of connected speech. Application of the CIU analysis to the conversation samples in this study was reported as reliable, with interrater reliability reaching a mean level of .88. Results indicated that although persons with aphasia spoke with a higher %CIU rate in conversational samples than in connected speech, their performance in the conversation samples could be predicted from their connected speech performance.

These findings seem to bolster the applicability of the CIU analysis as a means of measuring the communicative informativeness and efficiency of everyday language use. However, Doyle et al. suggest that additional research is needed to support this conclusion. By experimental design, the conversation samples in their study were obtained with *structured elicitation tasks*. Conversation was *elicited* as subjects were told when, where, how long, and what to talk about with greater (topic constrained) or lesser (topic open) specificity. Thus, the experimental constraints used to obtain conversation samples in this study preclude the generalizability of their findings to other conversational sampling conditions.

A question not addressed in research on the CIU analysis is whether it may be used to measure the communicative informativeness and efficiency of naturally occurring conversation. As Doyle et al. note (1995), a greater understanding of its applicability to authentic communicative interactions is needed if the %CIU is to be established as an ecologically valid tool for measuring communicative informativeness and efficiency of everyday language ability of persons with aphasia.

This study was designed as an initial effort toward this end. The purpose was to answer two basic questions: (1) Can the CIU analysis described by Nicholas and Brookshire (1993) for measuring communicative informativeness and efficiency of connected speech be applied reliably to the naturally occurring conversation of a person with aphasia? and (2) If it can be reliably applied, is the communicative informativeness and efficiency it measures a stable feature of the conversation of a person with aphasia across time and across conversational settings?

Methodological Issues

In the design of this study, several methodological issues were considered that related to research design and determination of reliability and stability of the CIU measure. The rationale for methodological decisions relating to these issues follows.

Research Design

In both of their articles examining the stability of their measures of informativeness and efficiency (Brookshire & Nicholas, 1994; Nicholas & Brookshire, 1993) and in Doyle et al.'s (1995) study, group designs were used. As Brookshire and Nicholas (1993) note, "Results based on group averages…do not speak directly of the test-retest stability of individual subjects' scores" (p. 402). They point out that clinical application of the CIU analysis requires individual assessment of reliability and stability of measures. Because of the importance of individual performance in clinical application, the conversation of a single subject was used for CIU analysis in this study. The limitation of generalizability of findings associated with study of a single case was accepted as compatible with the design of this study as an initial inquiry of the application of the CIU analysis to naturally occurring conversation.

Reliability

In all the studies utilizing the %CIU to date, each has had a primary rater score the entire discourse sample and then rescore a percentage of that sample to establish intrarater reliability. Interrater reliability was established in these studies by having two speech-language pathologists score a percentage of the discourse sample for comparison with the scores by the primary rater. This precedent was followed in the current study. Nicholas and Brookshire (1993) also report that their procedure included a chance for the raters to sit down and compare notes on how they were approaching their task after having scored a percentage of the connected speech sample. Doyle et al. (1995) do not report inclusion of such practice sessions in their study of structured conversational sampling. As the purpose of this study was to determine the reliability of the rule-based system for discourse for which it was not designed, any discussion that would potentially clear up ambiguity arising from differing interpretations of those rules could introduce a confounding element into this study. A test of reliability under these conditions would not be a test of the rules *as published*. Rather, it would be a test of the rules as published *when augmented by the unwritten rules agreed upon through a discussion between raters*. On the other hand, not including such opportunity could potentially detract from the clinical meaningfulness of this study, as it is common for clinicians to familiarize and even practice assessment tools before administration. Because of these contrasting concerns, the decision was made not to include any formal opportunity for practice or discussion but to provide raters with instructions to familiarize themselves with the rules and to take as much time as needed to complete the task. This provision allowed for raters to independently practice rule application. More specifically, no time constraints were placed on the raters' completion of their scoring of conversations (see Instructions 7 and 8 in the Appendix). The elapsed time for completion of scoring of all three raters was monitored and extended for as long as 9 hours. The length of time taken to complete the task as well as the extent and complexity of rules (see Appendix B, Nicholas & Brookshire, 1993, pp. 348-350) serve as indirect indices of the extent to which raters incorporated familiarization and independent practice as part of their participation in this study.

An additional issue related to the question of reliability is determination of the degree of rigor that is required before a measure is said to be reliable. There is currently no universally accepted standard for making this decision. (An extensive discussion of reliability standards is beyond the scope of this article. The interested reader is referred to Pedhazur & Schmelkin, 1991 and/or Seibel, 1968). There are guidelines in the literature, however, that suggest that relatively low reliabilities are acceptable for "ground breaking" research, that higher reliabilities are needed to determine differences between groups, and that only *very high reliabilities [can] be used to make decisions about individuals* (Pedhazur & Schmelkin, 1991).

The CIU analysis being examined in the current study is proposed to be both a research and a clinical measure. Important to clinical application specifically is that it is designed to be used in making clinical decisions regarding patient progress, continued treatment, and/or treatment protocols. Because clinical decisions like these *affect individuals' lives*, clinical measures must meet high reliability standards. The reliability of 90% found for the %CIU in Nicholas and Brookshire's (1993) original study and Doyle et al.'s report of 88% certainly are adequate for clinical decision making. However, because the current study applies the rule-based system to naturally occurring language phenomena for which it was not designed, it would be unrealistic to demand equally high reliability. On the other hand, a recent study by Crockford and Lesser (1994) of other measures of the everyday communication ability of persons with aphasia suggests that reliability in the low seventies is too low for clinical application. These authors examined the use of several clinical tools to validate efficacy of treatment. Included in their examination was the Communicative Effectiveness Index (CETI) (Lomas et al., 1989). The conduct of their study involved the administration of CETI by more than one scorer. Results indicated that the CETI had the least clinical utility for measuring change when compared to scores derived from other measures. The authors note that Lomas et al. report an interrater reliability of .73 (with a confidence interval of 95%) and suggest that this reliability level may be

insufficient to support the clinical application of this tool. Given the findings of Crockford and Lesser, which suggest that 73% is too low and the expectation that the 90% reliability obtained in other CIU studies may be too high, reliability scores greater than 80% were considered acceptable for the purposes of this study.

Method

Subject

Ed, a 50-year-old right-handed male, 6 years post stroke with history of a single left hemisphere CVA with residual moderate aphasia and mild right hemiparesis, served as the subject for this study. He was selected from among members of the local stroke club who volunteered for participation in research because of his similarity with the subject selection criteria of Nicholas and Brookshire (1993). Specifically, his Porch Index of Communicative Ability (PICA; Porch, 1981) percentile score was 62, which conformed to the aphasia severity of Nicholas and Brookshire's "average" subject (SPICA %ile average equaled 63.7). His Western Aphasia Battery (WAB; Kertesz, 1982) Aphasia Quotient was 46.6, with a WAB classification of conduction aphasia. His speech was intelligible, and he had sufficient language ability to actively participate in conversation. Ed was able to understand what was being said to him relatively well and was generally able to follow the topic of conversation. However, at times when comprehension was dependent on a single word or when the message was too long or linguistically complex, he would misunderstand. Although he was able to use many different forms of language and could almost always get his main idea across, his spontaneous speech was replete with instances of word-finding difficulty.

Conversational partners included (in all conversations and settings) his wife of 28 years, M, as well as (in multiparty conversations) the first author, MO, and her graduate research assistant, MG. All three conversational participants were ordinary speakers with no history of illness, disease, or deficit.

Data Collection

Conversations used in this study were collected as a part of a larger, on-going study of the conversational strategies of individuals with aphasia and their spouses (Oelschlaeger, 1999; Oelschlaeger & Damico, 1998a, 1998b).

To ensure authenticity, conversation collection procedures were conducted according to the commonly accepted protocol for the study of naturally occurring conversation (e.g., Atkinson & Heritage, 1994; Levinson, 1983). These procedures are well established in research on the naturally occurring conversation of ordinary speakers (e.g., Goodwin, 1987; Goodwin & Goodwin, 1986), persons with Alzheimer's disease (Orange, Lubinski, & Higginbotham, 1996), and persons with aphasia (Goodwin, 1995; Oelschlaeger, 1999; Oelschlaeger & Damico, 1998a, 1998b; Simmons-Mackie & Damico, 1996). Procedures included the videorecording of naturally occurring conversations over an extended period of time in a familiar setting. An 8-mm video camera was left in the subject's home for approximately 6 weeks. The couple was taught how to use the equipment and told to use it to record their conversations at their own discretion. No further direction was given as to when, where, how often, or in what manner of conversations were to be recorded. This resulted in the capturing of 5 conversations between the subject and his wife, recorded outdoors on their backyard patio and totaling 149 minutes of conversation. In addition to these dyadic conversations, 3 multiparty conversations were recorded in the couple's home. Conversational partners included—in addition to Ed and M—MO (two sessions) and MG (in all three conversations). These conversations took place with participants seated around the couple's dining room table and were allowed to develop naturally with no preagreed topic, length, or procedural arrangements. These conversations resulted in 117 minutes of taped discourse. In total, 266 minutes (4 hours and 26 minutes) of videorecorded conversation were obtained.

Videotaped conversations were viewed and transcribed by trained research assistants. Conversations were then reviewed many times by the authors while reading from these transcripts. Discrepancies between the video and the initial transcripts were resolved by consensus and corrections made. The transcript and videotapes were used to segment the conversations into turns (referred to as Turns at Talk, TAT, in this report) in keeping with the standard practice in the study of naturally occurring conversation (see, for example, Levinson, 1983; Orange et al., 1996; Sacks, Schegloff, & Jefferson, 1974).

Data Analysis

Conversation Participation

Information about Ed's participation in the eight conversations is presented in Table 1. As noted in Table 1, Ed took a total number of 1333 conversational turns and spoke a total of 8113 words in these conversations, an average of 6 words per turn. Of his turns 62% (830) were multiword utterances, an average of 104 per conversation. These findings indicate that he was an active participant in the conversations. More importantly, the number of conversations and the magnitude and the nature of his participation demonstrate how well these conversations represent Ed's naturally occurring

Table 1. Parameters of Ed's participation in conversation.

Conversation	Turns at Talk	Words per conversation	Average # words Turn at Talk**	Multiword Turns at Talk	Time talking***
A* 42 min	211	1595	8	124	17 min
B 28 min	212	1208	6	150	12 min
C* 31 min	169	1041	6	104	13 min
D 35 min	186	749	4	112	8 min
E* 44 min	184	1558	9	105	15 min
F 19 min	64	272	4	35	4 min
G 31 min	122	438	4	65	4 min
H 36 min	185	1252	7	135	16 min
Total: 266 min	Total : 1333	Total: 8113	Average: 6	Total: 830	Total : 89 min

*Multiparty conversations (All others were two-party conversations.)
**Rounded off to whole number
***Rounded to nearest minute

everyday language and, as such, offer objective support for authenticity of conversations.

Timing of Conversational Participation

Ed's Turns at Talk (TAT) in conversations were timed to provide temporal data needed for the efficiency calculations of the CIU analysis. However, because conversations are dialogic and naturally contain extended periods of time when no one is talking, the method used by Nicholas and Brookshire (1993) required modification. It was apparent that simply subtracting all contributions for participants other than the subject from the total time would grossly overestimate the amount of time Ed took to say what he had to say. For this reason, a method for timing Ed's talk was used that eliminated extended silences from the data set. Specifically, Ed's TAT were timed using a hand-held stopwatch. Pauses that occurred within the boundaries of his TAT were included in the total time for that TAT. Pauses that occurred outside TAT boundaries were not included. Short, single-word TAT or extremely short, multiword TAT (i.e., "Oh yeah") were given a standard time score of .25 seconds, as this was the lower time limit for which the watch being used could be manually started and then stopped. Ten percent of Ed's TAT were retimed as an intraexaminer reliability measure. Point-to-point reliability of TAT times was 90.19%, with a margin of error of ±.10 of a second.

Reliability of Word and CIU Counts

A complete data set consisting of word count, CIU count, and the derived %CIU was obtained from the application of the rule-based system for CIU analysis to Ed's talk in all conversations by the primary rater (the second author). Intrarater reliability calculations were generated 4 weeks after initial counts were made. At this time, the primary rater rescored Ed's talk in a randomly selected conversation from each conversational setting (one dyadic, one multiparty). Ed's TAT in these two conversations represent approximately 25% of the TAT available for analysis.

Data for interrater reliability were generated by having two speech-language pathologists (Raters One and Two) score Ed's TAT in the two randomly selected conversations used in the intrarater reliability calculations. Both speech-language pathologists serving as independent raters in this study were ASHA-certified and had been professionally involved in the diagnosis and treatment of persons with aphasia for over 3 years. Both were currently employed, one in a rehabilitation setting and the other as a university clinical supervisor of graduate students clinically involved with neurologically impaired adults. Neither Rater One or Two had any previous interaction with the subject.

Raters were told they were participating in a study examining the applicability of a standardized measure of conversation. They were given a packet containing a full transcript (i.e., transcripts included the conversational turns of every participant) of each conversation. They were also given a copy of the rules for making word and CIU counts and a set of written instructions specific to the procedures of the study (see Appendix). Scoring was done independently by each rater in a controlled environment.

Both interrater and intrarater reliability for word and CIU counts of Ed's TAT were calculated by dividing the number of agreements by the sum of the agreements and disagreements and multiplying by 100. A point-to-point reliability calculation was done. These calculations were made on a TAT-by-TAT basis, rather than a word-by-word basis, as word count was one of the measures being studied. Specifically, the number of CIUs and/or

words for each turn at talk for each rater was totaled. These totals were then compared for agreement across raters. Agreement was considered to exist when equal scores were given by raters for the total number of words or CIUs in a particular TAT.

Results

Reliability of Word Count, CIU Count, and %CIU

Intrarater and interrater reliability for all measures, raters, and conversations is reported in Table 2. Data are reported first as an average reliability score for the entire data sample. Data on reliability found in each conversational setting are then reported so that comparison across conversational settings can be made.

Word Count Reliability

As seen in Table 2, both the intrarater and interrater reliability of word count were found to be high (greater than 90%) for all raters across both conversation settings.

CIU Count Reliability

Conversely, both intrarater and interrater reliability for CIU counts was poor (less than 73% for intrarater and less than 56% for interrater). Conversational setting had no apparent impact on these scores, as scores from each setting fell within five percentage points of each other.

%CIU Reliability

The %CIU is a calculated measure combining word count and CIU count. Its reliability is, therefore, affected by the reliability of the measures used in its calculation. Reliability for %CIU was calculated on an agreement/disagreement basis. Intrarater reliability for %CIU was higher (73%) than that for interrater reliability (never greater than 55%). Both fell short of an acceptable range of 80%, the established criterion level for the purposes of this study. Though the variation between scores in the different conversational settings was greater for %CIU than for the other measures, no pattern emerged that would indicate a consistent influence of conversational setting on the reliability of the measure.

During analysis of the reliability data, it was noted that within individual TAT, different types of disagreement occurred between the different raters. For example, two of the raters may have agreed on the number of words in a particular TAT, but disagreed on the number of CIUs. On this same TAT, the third rater may have disagreed with the other raters' word count, but agreed with the CIU count of one of the raters. To get a more accurate picture of the degree to which this was occurring, word and CIU counts were analyzed to examine what percentage of TAT were judged to have equal numbers of *both* word and CIU counts by all three raters. This analysis found a consensus among raters for only 40.65% of TAT in the dyadic conversation and for 37.43% of TAT in the multiparty conversation. This resulted in a consensus for both word and CIU count for only 39% of TAT in the overall reliability sample.

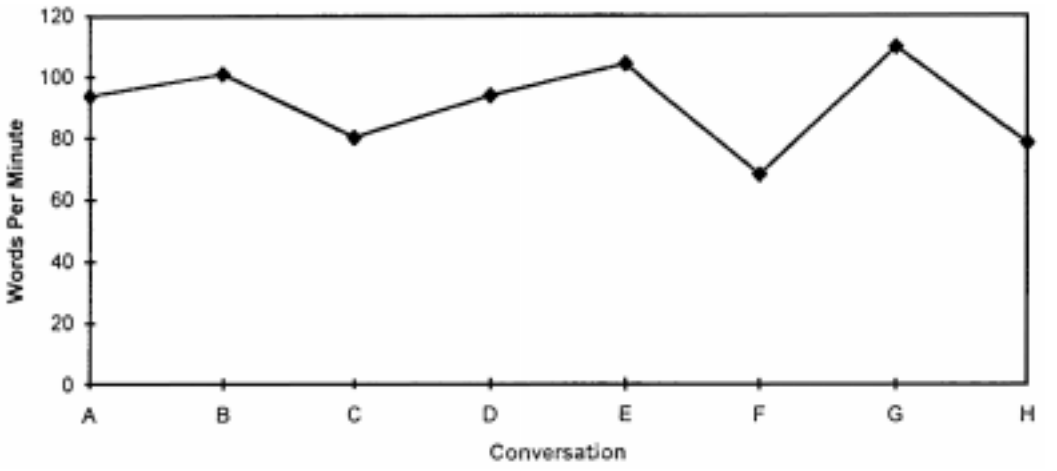
Table 2. Intrarater and interrater reliability for all measures by rater.

Rater	Word count (%)	CIU count (%)	%CIU
Intrarater reliability			
Author			
Total sample	98	72	73
Dyadic conversation	99.1	71.96	72.43
Multiparty conversation	97.8	72.63	72.63
Interrater reliability			
Author to Rater One			
Total sample	98	56	55
Dyadic conversation	96.26	57.28	57.01
Multiparty conversation	99.12	54.19	53.07
Author to Rater Two			
Total sample	91	56	55
Dyadic conversation	91.59	53.27	52.8
Multiparty conversation	90.05	58.66	57.54
Rater One to Rater Two			
Total sample	91	52	51
Dyadic conversation	90.19	52.34	51.4
Multiparty conversation	91.06	50.84	50.28
Average across all raters and both conversations	93	54	54

Stability of Measures

Assessment of the stability of measures was designed according to the recommendations of Brookshire and Nicholas (1994). That is, Ed's TAT were segmented into 300 word samples to allow correlational comparisons (mean, standard deviation, variance, and Pearson Product Moment) of both speech rate and %CIU to be made across time and conversational settings. As reliability of the measures was not established, stability calculations were not made. Nonetheless, it was possible to determine the stability of speech rate because these calculations had been made in preparation for determination of the stability of the %CIU measure. Results are graphically presented in Figure 1. Speech

Figure 1. Ed's average speech rate per conversation.



rate was noted to be relatively stable, averaging 91 words per minute with a range of 68 to 110 words per minute.

Post Hoc Analysis

The results reported above paint a rather complicated picture. Although it is clear that reliable scores could not be obtained for both CIU counts and %CIU when the rule-based system for CIU analysis was applied to the naturally occurring conversation of a person with aphasia, the reasons for the lack of reliability are not apparent in the data. As Porch (personal communication, 1997) observes, there are two primary causes for poor reliability: incompetence of the raters or insufficiency of the scoring system. Given the desire in clinical aphasiology to develop a way to capture the communicative informativeness and efficiency of everyday language and the high reliability reported in all the other published studies using the CIU analysis, additional analyses were performed. Specifically, post hoc analyses were conducted to explore whether the raters were not competent raters or the rule system was not sufficient in some way and, therefore, did not allow raters to make reliable judgments when analyzing the naturally occurring conversation of an individual with aphasia.

Competence of the Raters

The question of the competence of the raters was approached in two ways. The first involved analysis of the intrarater reliability of the independent raters. The second involved analysis of the accuracy of application (e.g., human error) of the rules by all three raters.

Intrarater reliability of raters. To determine the consistency with which these independent raters applied the rules for making word and CIU counts, the two speech-language pathologists were asked to rescore a portion of the conversations they had scored originally. They rescored a randomly selected portion of TAT equal to approximately 20% of the original sample approximately 6 weeks after their initial scoring sessions. Intrarater reliability was then calculated for each rater using the formula and methods described above. Results of this analysis are presented in Table 3. Data are reported first as an average reliability score for the entire data sample. Data on reliability found in each conversational setting are then reported so that comparison across conversational settings can be made.

As noted in Table 3, intrarater reliability of the independent raters for word count was good, averaging 86%. This is considered to be within a range acceptable for most purposes. However, the intrarater reliability scores for both CIU count and %CIU were not within an acceptable range, reaching only 61% and 57% for Rater One and 54% and 53% for Rater Two respectively. Again no consistent pattern emerged regarding conversational setting in relation to these scores. These findings, combined with those of the primary analysis (see Table 2), show inconsistency in application of the CIU analysis rules within and across all raters.

Accuracy of rule application. The second approach

Table 3. Intrarater reliability for independent raters.

Rater	Word count (%)	CIU count (%)	%CIU
Rater One			
Total sample	85	61	57
Dyadic conversation	90.57	56.6	54.72
Multiparty conversation	79.41	64.71	58.82
Rater Two			
Total sample	87	54	53
Dyadic conversation	88.68	58.49	56.6
Multiparty conversation	85.29	50	50

to investigating rater competency involved examining the accuracy with which all raters applied the rule-based system for CIU analysis. The rules for scoring and counting words and correct information units (CIUs) are published in Appendix B in Nicholas and Brookshire's 1993 article (pp. 348–350). The rules are extensive, text-based, and complex as they include inclusion and exclusion criteria to be applied to scoring of speech as a correct information unit.

To determine accuracy of rule application, all TAT for which there was disagreement between raters were identified by the second author. In total, 350 discrepancies were noted. Then, the discrepancies were categorized as conforming to the rules or as a misapplication of an unambiguous rule (i.e., human error). For example, in the dyadic conversation, one rater determined that one of Ed's TAT, "And Norma's is two," had 4 CIUs, counting *And* at the beginning of the TAT as a CIU, whereas the other raters counted only 3 CIUs, excluding *And* from their count. As the word *and* is excluded from all CIU counts according to Nicholas and Brookshire's (1993) Rule 2.20 ("*And* is never counted as a correct information unit…," p. 349), the disagreement between raters on that TAT was attributed to human error. Evaluation of the reliability of this categorization involved the independent categorization of the 350 discrepancies by the first author. The proportion of agreement was 96%, using the formula of agreements divided by agreements plus disagreements times 100.

Results of this analysis were that 99 of the 350 discrepancies were attributable to misapplication of an unambiguous rule to the TAT by one or more of the raters. These 99 instances accounted for just over one quarter (28%) of the disagreements between raters. An easy majority of these errors fell into two categories: counting *and* as a CIU and counting vague words such as *thing* or *stuff* as CIUs. Another commonly occurring error involved one rater's misapplication of the rules that define whether or not a conjunction is a CIU. (This rater excluded nearly all coordinating conjunctions from her count, overgeneralizing a rule that excludes conjunctions used as filler.)

Clearly, human error influenced the reliability noted in this study. To determine the extent of this influence, these 99 disagreements were eliminated from the data set, and interrater reliability was recalculated. The resulting reliability was 62.19, which again was below the criterion-level agreement set in this study and indicated that other factors were affecting reliability.

Sufficiency of the Rule-Based System

To examine the sufficiency of the rule system for application to naturally occurring conversation, an analysis was performed on the remaining 251 (72%) TAT where raters disagreed and where the disagreement could not readily be attributed to a misapplication of a rule. For example, in the multiparty conversation, one of Ed's TAT, "I would say, Italian was Italian. Go, there were some Mexicans, but maybe not many," was determined to have 15 CIUs by one rater who excluded only the word *Go* in accordance with Rule 2.13, which excludes "dead end, false starts or revisions in which the speaker begins an utterance but either revises it or leaves it uncompleted and uninformative with regard to the picture(s) or topic" (Nicholas & Brookshire, 1993, p. 348). Another rater credited Ed with only 9 CIUs in this same utterance. She excluded from her count "I would say" and "Italian was" in addition to the word *Go*. An examination of each of the disputed pieces of this TAT revealed at least two rules that could be applied in order to determine if they were correct and informative. For example, the phrase "I would say" could easily be interpreted as excluded according to rule 2.21. This rule excludes "commentary on the task and lead-in phrases that do not give information about the picture(s) or task and are not necessary for the grammatical completeness of the statement" (Nicholas & Brookshire, 1993, p. 349). If, however, a rater interpreted the phrase "I would say" to be equivalent to the phrase "I would estimate" or similarly "It is my opinion that," then the phrase could not be excluded by the rule. In this case the phrase would be thought to convey relevant, correct information to the listener and would, therefore, be considered as correct in context (Rule 3.11) and counted as 3 CIUs. Because both positions can be argued reasonably from the rules, the disagreement between raters was attributed to rule insufficiency rather than human error on the part of the raters, and the linguistic and conversational context in which the disagreement occurred was noted. When this had been done for each of the disagreements in both conversations, these various linguistic and interactional features were classified in order to examine patterns or trends in the data. This led to the identification of two broad categories into which these 251 discrepancies could be placed. The first broad category related to the raters' interpretations of informativeness; the second related to the absence of rules covering specific interactive language use found in the conversations.

Interpretations of informativeness. Of these two categories, interpretations of informativeness was the largest, accounting for 159 (74%) of the 251 disagreements that were not accounted for by human error. In this category, disagreements between raters seemed to stem from the subjectivity involved in applying the central principle of correctness used in the CIU analysis to the complex interactive environment found in conversation. The scoring rules indicate that words and phrases need to be "correct in context," even if not grammatically well

formed, if they are to be counted as a correct information unit. This presupposes that a standard of correctness is discernable in the sample being scored. Nicholas and Brookshire (1993) selected elicitation stimuli to constrain the types and topics of the connected speech they sampled. However, in natural conversation, the types of discourse present and topics covered are collaboratively determined by participants (Sacks, 1992). For this reason, there is a much greater ambiguity from moment to moment regarding exactly what direction the conversation will take. This inherent ambiguity affects the determination of correctness by a rater attempting to analyze the conversation, particularly when one or more of the conversational participants is a disordered speaker. In the current study, disagreements resulting from these factors were found either in TAT that contained a significant amount of revision and/or grammatical error or in TAT with a degree of ambiguity sufficient to support multiple interpretations. Because the full transcript was provided, each rater had information available to them from the conversation context (e.g., turns at talk that were antecedent and subsequent to Ed's turn) to use in making a determination of the informativeness of Ed's talk. However, because the rule system was designed to measure single-speaker discourse, it does not provide guidelines on how this interactive information may be used. Therefore, each rater approached the job of determining just which words in Ed's TAT were correct and informative differently. As a result, each rater's count of which words in a particular TAT conformed to the definition of a CIU was also different.

An example from the data will help to clarify how this occurred. In the dyadic conversation, Ed said, "Now we, now we're clean all the things that uh, vacuuming, now we, all we have to do is this." One rater excluded almost this entire TAT from her CIU count, giving Ed credit for "…we're…vacuuming…" and scoring his TAT at 3 CIUs. Another rater excluded much less of this TAT, giving Ed credit for "…now we're clean…vacuuming…all we have to do is this" and scoring his TAT at 12 CIUs. The third rater excluded still less of this statement giving Ed credit for "…now we're clean all the…vacuuming, now… all we have to do is this" and scoring his TAT at 15 CIUs. Although it is impossible to say exactly what logic each rater used to decide which of the words in this 20-word TAT provided correct information to listeners, it is clear that their interpretations of Ed's message were slightly different. The first rater seems to have decided that Ed provided the information "we were vacuuming." The second rater interpreted Ed's message differently, deciding it contained more correct information, something to the effect that "we're clean, have done the vacuuming, and all we have left to do is this." The third rater's interpretation of the message was similar to the second rater's, but allowed for more of it to be considered informative. In this case the message could be construed to have been something along the lines of "Now that we are clean, having done *all* the vacuuming, all we have left to do is this." When each of the raters had decided which words conveyed correct information in the context of each rater's particular interpretation of the message, the rest of the TAT was excluded from their CIU count. The determination of what was "correct in context" in this instance was quite different for each rater. As a result their CIU counts were different.

Absence of rules. The second broad category of disagreement between the raters, accounting for 92 (just over 26%) of the 251 disagreements between raters not attributable to human error, resulted from absence of rules for defining CIUs for the unique interactive language use that occurs in conversation and not in single-speaker connected speech. Data analysis identified six subcategorizes of TAT with specific linguistic and contextual features not addressed in the rule-based system for CIU analysis. An example of each subcategory is presented in Table 4.

As displayed in Table 4, the subcategories relating to absence of rules were (a) *yeah* and equivalent affirmatives (such as *right* and *uh huh*) as an answer to a direct question, (b) *yeah* and other interactive tools (such as *uh huh* and *hmm*) used by Ed to acknowledge or encourage the speaker, (c) automatic speech such as social greetings or idiomatic phrases (*what the hell* and *maybe, maybe not*, for example), (d) direct questions, (e) repetition of another speaker's remark to show agreement or acknowledgment, and (f) incomplete utterances resulting from interruption.

Discussion

The purpose of this study was to address two primary questions: (1) Can the rule-based system for CIU analysis designed by Nicholas and Brookshire (1993) for measuring communicative informativeness and efficiency of connected speech be applied reliably to the naturally occurring conversation of a person with aphasia? and (2) Is the communicative informativeness and efficiency it measures a stable feature across time and across conversational contexts? The answer to the first of these questions was that no, as a whole, the CIU analysis could not be applied reliably to the conversation of the single subject of this study. The answer to the second question was precluded by the reliability findings, although speech rate was noted to be stable across conversations.

Two competing possible explanations for the low reliability noted in this study were identified. One was

Table 4. Subcategories and examples of absence of rules.

Subcategory	Example		
	Turn No.	Speaker[a]	Utterance
1. *Yeah* as an answer to a direct question.	274	MO	You mean for your person? Or for
	275	Ed	/ / Yeah.
2. *Yeah* as an interactive tool.	63	M	It's a pretty day.
	64	Ed	Yeah.
3. Automatics.	75	M	The wall doesn't look good, does it?
	76	Ed	No, but what the hell.
4. Questions.	43	M	There it goes. Two-ton. Kitty, kitty, kitty.
	44	Ed	Him?
5. Repetition of other speakers.	383	MG	Now that's not all the bleachers.
	384	M	That's it.
	385	Ed	That's it.
6. Incomplete utterance due to interruption by other speaker.	96	Ed	OK. Approval and everything. But the card is the
	97	M	/ / the American Express or the credit card.

[a]MO = first author. M = wife. MG = research assistant.

that the raters were not up to the task. The other was that there were insufficiencies in the rule-based system which precluded its reliable application to naturally occurring conversation. The majority of evidence indicated that the observed low reliability stemmed primarily from rule insufficiency rather than rater competence. Direct evidence relating to rater competence was noted in examination of the reliability of word counts. Specifically, the high reliability for word counts (intrarater and interrater scores of 93% and above) indicated that *when given a sufficiently clear set of rules,* raters made clear and consistent judgments regarding the fit of the data to those rules. Less direct but supportive evidence of rater competence was gained in analysis of errors resulting in discrepancies between raters. Specifically, human error did occur but accounted for only 28% of the total 350 discrepancies. Also, even if human errors were totally eliminated from the data set, the overall interrater reliability for CIU of 69.12 would still fall below the criterion of 80% set for the purposes of this study. Finally, support for identifying raters as competent (i.e., capable of producing reliable scores) was gained from the dissimilarity between CIU reliability and word-count reliability. If the raters' ability to make consistent judgments were the only factor affecting reliability, then it would be expected that reliability scores for the two primary measures (word and CIU counts that are used to calculate %CIU) would be at least similar. This was most definitely not the case; CIU reliability scores fell as much as 40 percentage points below Word Count reliability scores.

Even though the majority of evidence suggested rater competence, human error did occur; and the issue of raters' practicing scoring rules before applying the CIU analysis to any language sample warrants mention in this regard. The procedures for determining reliability in this study followed those used in Nicholas and Brookshire's (1993) initial report of the CIU analysis in all aspects but one: No opportunity for discussion between raters relating to the rule system was provided. However, an opportunity for "on the job training" of raters was provided. Although reliability for CIU was below acceptable levels for all raters, the primary rater (second author) had higher intrarater reliability scores than the other two raters. Given that the primary rater had greater exposure to the CIU rules through his participation in the design and conduct of this study and the scoring of all eight conversations, the possibility arises that these experiences constituted a greater "self-training" opportunity. In addition, the existence of 99 discrepancies between raters that were readily attributable to human error suggests that at least some improvement in reliability (albeit not enough to meet the reliability criteria of this study) would be gained if raters practiced the CIU administration before applying it to naturally occurring conversation. Although beyond the scope of this study, the importance of practice merits consideration in future research.

The evidence that insufficiencies in the rule system played a primary role in the poor reliability of the data collected comes from the finding that a full 72% (251) of the 350 disagreements between raters were attributable

to differing, but competent, interpretations of the rules. It follows, therefore, that the rule-based system designed for connected speech does not provide the guidance needed by raters if they are to make reliable judgments in the complex interactive environment of conversation. This conclusion is further supported by the fact that 26% (92) of the 251 disagreements occurred in conversational contexts not addressed by the rules.

This result is not surprising if it is remembered that the rules for the CIU were designed to be used on data collected from single-speaker discourse samples (connected speech) and therefore did not need to address the interactive issues found in conversation. As well, in their design of the measure, Nicholas and Brookshire (1993) used discourse elicitation tasks designed to constrain the possible range of "correct," and therefore informative, responses. As a result, the rule-based system did not need to address the issue of how a reliable framework for correctness would be established by raters as they scored samples. Raters were expected to already have a stable framework available to them when they used the system (provided by a picture sequence, for example). A quite different situation prevails when raters are asked to apply the CIU analysis to naturally occurring conversation. Because no predetermined framework of correctness exists in conversation, raters in this study had to develop their own. Presumably, they did so by capitalizing on the information resource provided by the conversational context. However, without rule guidance on the use of contextual information, differing interpretations of the informativeness of Ed's talk occurred. That is, because no predetermined framework of correctness exists in conversation, raters in the current study were left on their own, without guidance from the rules, when determining the correctness of a particular word in the sample. As a result, scores differed for different raters as they brought their varied resources and experience to the task at hand.

Reconciliation of Doyle et al.'s (1995) report of high reliability (.88) when the CIU analysis was applied to conversation samples with the results of this study is less straightforward, and several explanatory possibilities may be considered. One possibility is that we were successful, as intended, in investigating the application of the CIU analysis to everyday language not previously studied (e.g., naturally occurring conversation). That is, connected speech differs from elicited conversation samples, which differ from naturally occurring conversation. Thus,application of the CIU analysis, designed only for connected speech, varies in reliability accordingly. However, the comparability of interrater reliability levels in Doyle et al. and Nicholas and Brookshire's studies (.88 and .90 respectively) and the dissimilarity of both with ours (.54) raises the question of whether Doyle et al.'s conversation samples were "more like" the connected speech of a single speaker than the interactive language displayed in the naturally occurring conversation of our subject. If so and the Doyle et al. study lacked the framework of correctness that influenced our results, that would account for differences in results between studies. A second possibility is that Doyle et al.'s conversation samples did not differ from ours but that the raters in their study had experiences that fostered the development of similar frameworks for judging correct information units that our raters did not have. In Doyle et al.'s study, two authors served as raters. It seems fair to assume that they had a priori knowledge of the conversational topic and some expectation of content for at least half of the conversation samples (e.g., those obtained in their "topic constrained" condition). Thus, when faced with having to score a unit of speech as correct or not correct, without guidance from the rules of the CIU analysis, they were able to draw on this knowledge to develop similar frameworks for making judgments. Because naturally occurring conversation is dynamically and interactively directed by participants, knowledge of topic and expected content was not available to our raters. Thus, they independently developed frameworks that were quite dissimilar. A third possibility for differences in findings relates to the primary design difference in studies. We used a single subject, whereas all previous studies used group designs. Even though our subject fell within the range of aphasia severity and type noted in Doyle et al.'s study, it is possible that his participation in conversation was more challenging to score than that of the "average" person with aphasia.

Clearly, additional research on the application of the CIU analysis to the naturally occurring conversation of other individuals with aphasia is needed—not just to support the generalizability of the results of this study but also to clarify differences in results across studies. This research should include a comparison of the reliability and stability of the performance of single individuals on the original CIU analysis protocol with their performance in naturally occurring conversation. The study of normal, ordinary speakers would also offer meaningful information about applicability of the CIU analysis to naturally occurring conversation.

Future research is also needed to extend the information provided by the current study. This would include any study that fostered the development of a rule system designed specifically to address the complexities of determining concepts of informativeness and correctness in the complex interactive environment of naturally occurring conversation. Findings of this study suggest that a better understanding of how people produce meaning in a conversation (particularly when one or more of the speakers is a disordered speaker) is central to success of any such system. With such an understanding, guidelines can be designed that provide raters with

an interpretive framework upon which they can base their judgments of correctness/informativeness. Even without this understanding, however, we can expect that a successful rule system would include rules for assessing informativeness based on the impact that the subjects' utterances have on their conversational partners. An utterance, for example, might be considered informative if it does not elicit any repair requests from the conversational partner; if it does elicit a repair request, it might be considered informative if that repair request is specific rather than general (i.e. "When did you do that?" as opposed to "What?"). The rules would need to include guidelines for incorporating the information provided by other speakers; for making decisions regarding idiomatic, social, and automatic phrases; and for handling speech that is used by participants to organize their conversational interaction. A successful system would need to include rules addressing question and answer sequences, interruptions, and other conversational elements.

Clinical application of the CIU analysis is certainly warranted for assessment of connected speech as described by Nicholas and Brookshire (1993) or the conversation conditions described by Doyle et al. (1995). However, this study suggests that capturing everyday language ability as exemplified by the naturally occurring conversation of a person with aphasia with the CIU rule-based system as currently designed remains a challenge to the field.

Acknowledgments

This study was conducted as a master's thesis by the second author that was directed by the first author. We would like to express our appreciation to Bob Brookshire, Bob Marshall, and Associate Editor Shari Baum for their objectivity in reviewing this work and for their meaningful comments.

References

Atkinson, J. M., & Heritage, J. (1994). *Structures of social action: Studies in conversation analysis*. Cambridge, U.K.: Cambridge University Press.

Blomert, L., Koster, C., Van Mier, H. & Kean, M-L. (1987). Verbal communication abilities of aphasic patients: The everyday language test. *Aphasiology, 1,* 463–474.

Brookshire, R. H., & Nicholas, L. E. (1995). Performance deviations in the connected speech of adults with no brain damage and adults with aphasia. *American Journal of Speech-Language Pathology, 4*(4), 118–123.

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and the test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37,* 399–407.

Crockford, C., & Lesser, R. (1994). Assessing functional communication in aphasia: Clinical utility and time demands of three methods. *European Journal of Disorders of Communication, 29,* 165–182.

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *American Journal of Speech-Language Pathology, 4*(4), 130–134.

Doyle, P. J., Tsironas, D., Goda, A. J., & Kalinyak, M. (1996). The relationship between objective measures and listeners' judgments of the communicative informativeness of the connected discourse of adults with aphasia. *American Journal of Speech-Language Pathology, 5*(3), 53–60.

Goodwin, C. (1995). Co-constructing meaning in conversations with an aphasic man. *Research on Language and Social Interaction, 28,* 233–260.

Goodwin, C. (1987). Forgetfulness as an interactive resource. *Social Psychology Quarterly, 2,* 115–131.

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica, 62,* 51–72.

Goodglass, H., & Kaplan, E. (1972). *Boston Diagnostic Aphasia Examination.* Philadelphia: Lea and Febiger.

Helmsley, G., & Code, C. (1996). Interactions between recovery in aphasia, emotional and psychosocial factors in subjects with aphasia, their significant others and speech pathologists. *Disability and Rehabilitation, 18,* 567–584.

Holland, A. (1980). *Communicative Abilities in Daily Living.* Baltimore: University Park Press.

Kertesz, A. (1982). *The Western Aphasia Battery*. New York: Grune & Stratton.

Levinson, S. C. (1983). *Pragmatics.* Cambridge, U.K.: Cambridge University Press.

Lomas, J., Pickard, L., Bester, S., Elbard, H., Finlayson, A., & Zoghaib, C. (1989). The Communicative Effectiveness Index: Development and psychometric evaluation of a functional communication measure for adult aphasia. *Journal of Speech and Hearing Disorders, 54,* 113–124.

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research, 38,* 145–156.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36,* 338–350.

Oelschlaeger, M. L. (1999). Participation of a conversation partner in the word searches of a person with aphasia. *American Journal of Speech-Language Pathology, 8,* 62–71.

Oelschlaeger, M. L., & Damico, J. S. (1998a). Joint productions as a conversational strategy in aphasia. *Clinical Linguistics and Phonetics, 12,* 459–480.

Oelschlaeger, M. L., & Damico, J. S. (1998b). Spontaneous verbal repetition: A social strategy in aphasic conversation. *Aphasiology, 12,* 971–988.

Orange, J. B., Lubinski, R. B., & Higginbotham, D. J. (1996). Conversational repair by individuals with

dementia of the Alzheimer's type. *Journal of Speech and Hearing Research, 39,* 881–895.

Pedhazur, E. J., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach.* Hilldale, NJ: Lawrence Erlbaum Assoc.

Porch, B. E. (1981). *The Porch Index of Communicative Ability.* Palo Alto, CA: Consulting Psychologists Press.

Sacks, H. (1992). *Lectures on conversation.* Cambridge, U.K.: Basil Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50,* 361–382.

Seibel, D. W. (1968). Measurement of aptitude and achievement. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences* (pp. 261–314). Reading, MA: Addison-Wesley.

Simmons-Mackie, N. N., & Damico, J. S. (1996). The contribution of discourse markers to communicative competence in aphasia. *American Journal of Speech-Language Pathology, 5,* 37–43.

Received January 1, 1998

Accepted August 25, 1998

Contact author: Mary L. Oelschlaeger, PhD, Department of Speech Pathology and Audiology, Northern Arizona University, PO Box 15045, Flagstaff, AZ 86011. Email: mary.oelschlaeger@nua.edu

Appendix. Rater packet.

Instructions:

This study is an examination of a system for measuring the communicative informativeness and efficiency of connected speech of a person with aphasia. Your job will be to apply a set of rules (included in your packet) to the connected speech of one participant in two conversations. The rules will generate two kinds of information: a word count and a count of Correct Information Units (CIUs). You will count both words and CIUs from a transcript of the two conversations. Please follow these steps when recording your word and CIU counts.

1. Score only what is said by our subject with aphasia, Ed (E).
2. Record your count of words in the blank marked "W" at the end of each of Ed's turns.
3. Record your count of CIUs in the blank marked "CIU" at the end of each of Ed's turns.
4. Do not score any turn which appears in *italics.* These are included in order to supply context only.
5. Do not score any words appearing in parentheses. These are either best guesses at unintelligible utterances or editorial comments. Again these are provided in order to provide context.
6. The symbol "*" is used to indicate unintelligible utterances.
7. Please read the complete set of rules prior to beginning your scoring. There are rules for both words and CIUs.
8. I will be keeping track of how long this task takes. You have as long as you need in order to complete this task. If you are not finished by the end of the time we have arranged for today, we will schedule a time when scoring can be completed.
9. If you need anything from me while you are scoring, please feel free to ask.

Thank you so much for your help with this study, and good luck.

Rules for Word and CIU Count:

Raters were also given a photocopy of the rules from the original article on the %CIU: Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36,* 338–350. The rules may be found in Appendix B of their article on pages 348 to 350.