# Evaluating Discourse and Dialogue Coding Schemes

Richard Craggs*
University of Manchester

Mary McGee Wood*
University of Manchester

*Agreement statistics play an important role in the evaluation of coding schemes for discourse and dialogue. Unfortunately there is a lack of understanding regarding appropriate agreement measures and how their results should be interpreted. In this article we describe the role of agreement measures and argue that only chance-corrected measures that assume a common distribution of labels for all coders are suitable for measuring agreement in reliability studies. We then provide recommendations for how reliability should be inferred from the results of agreement statistics.*

Since Jean Carletta (1996) exposed computational linguists to the desirability of using chance-corrected agreement statistics to infer the reliability of data generated by applying coding schemes, there has been a general acceptance of their use within the field. However, there are prevailing misunderstandings concerning agreement statistics and the meaning of reliability.

Investigation of new dialogue types and genres has been shown to reveal new phenomena in dialogue that are ill suited to annotation by current methods and also new annotation schemes that are qualitatively different from those commonly used in dialogue analysis. Previously prescribed practices for evaluating coding schemes become less applicable as annotation schemes become more sophisticated. To compensate, we need a greater understanding of reliability statistics and how they should be interpreted. In this article we discuss the purpose of reliability testing, address certain misunderstandings, and make recommendations regarding the way in which coding schemes should be evaluated.

## 1. Agreement, Reliability, and Coding Schemes

After developing schemes for annotating discourse or dialogue, it is necessary to assess their suitability for the purpose for which they are designed. Although no statistical test can determine whether any form of annotation is worthwhile or how applications will benefit from it, we at least need to show that coders are capable of performing the annotation. This often means assessing reliability based on agreement between annotators applying the scheme. Agreement measures are discussed in detail in section 2.

Much of the confusion regarding which agreement measures to apply and how their results should be interpreted stems from a lack of understanding of what it means to

* School of Computer Science, University of Manchester, Manchester, M13 9PL, U.K.
E-mail: richard_craggs@yahoo.co.uk; mary_mcgee.wood@manchester.ac.uk.

assess reliability. For example, the coding manual for the Switchboard DAMSL dialogue act annotation scheme (Jurafsky, Shriberg, and Biasca 1997, page 2) states that kappa is used to "assess labelling accuracy," and Di Eugenio and Glass (2004) relate reliability to "the objectivity of decisions," whereas Carletta (1996) regards reliability as the degree to which we understand the judgments that annotators are asked to make. Although most researchers recognize that reporting agreement statistics is an important part of evaluating coding schemes, there is frequently a lack of understanding of what the figures actually *mean*.

The intended meaning of reliability should refer to the degree to which the data generated by coders applying a scheme can be relied upon. If we consider the coding process to involve mapping units of analysis onto categories, data are reliable if coders agree on the category onto which each unit should be mapped. The further from perfect agreement that coders stray, the less we can rely on the resulting annotation.

If data produced by applying a scheme are shown to be reliable, then we have established two important properties of those data:

1. The categories onto which the units are mapped are not inordinately dependent on the idiosyncratic judgments of any individual coder.
2. There is a shared understanding of the meaning of the categories and how data are mapped onto them.

The first of these is important for ensuring the reproducibility of the coding. To be able to trust the analysis of annotated corpora, we need to be confident that the categorization of the units of data is not dependent on which individual performed the annotation. The second governs the value of data resulting from the coding process. For an annotated corpus or the analysis thereof to be valuable, the phenomenon being annotated must represent some notion in which we can enjoy a shared understanding.

## 2. Agreement Measures

There are many ways in which the level of agreement between coders can be evaluated, and the choice of which to apply in order to assess reliability is the source of much confusion. An appropriate statistic for this purpose must measure agreement as a function of the coding process and not of the coders, data, or categories. Only if the results of a test are solely dependent on the degree to which there is a shared understanding of how the phenomena to be described are mapped to the given categories can we infer the reliability of the resulting data. Some agreement measures do not behave in this manner and are therefore unsuitable for evaluating reliability.

A great deal of importance is placed on domain specificity in discourse and dialogue studies and as such, researchers are often encouraged to evaluate schemes using corpora from more than one domain. Concerning agreement, this encouragement is misplaced. Since an appropriate agreement measure is a function of only the coding process, if the original agreement test is performed in a scientifically sound manner, little more can be proved by applying it again to different data. Any differences in the results between corpora are a function of the variance between samples and not of the reliability of the coding scheme.

Di Eugenio and Glass (2004) identify three general classes of agreement statistics and suggest that all three should be used in conjunction in order to accurately evaluate coding schemes. However, this suggestion is founded on some misunderstandings of the role of agreement measure in reliability studies. We shall now rectify these and conclude that only one class of agreement measure is suitable.

### 2.1 Percentage Agreement

The first of the recommended agreement tests, **percentage agreement,** measures the proportion of agreements between coders. This is an unsuitable measure for inferring reliability, and it was the use of this measure that prompted Carletta (1996) to recommend chance-corrected measures.

Percentage agreement is inappropriate for inferring reliability because it excludes any notion of the level of agreement that we could expect to achieve by chance. Reliability should be inferred by locating the achieved level of agreement on a scale between the best possible (coders agree perfectly) and the worst possible (coders do not understand or cannot perform the mapping and behave randomly). Without any indication of the agreement that coders would achieve by behaving randomly, any deviation from perfect agreement is uninterpretable (Krippendorff 2004b).

The justification given for using percentage agreement is that it does not suffer from what Di Eugenio and Glass (2004) referred to as the "prevalence problem." *Prevalence* refers to the unequal distribution of label use by coders. For example, Table 1 shows an example taken from Di Eugenio and Glass (2004) showing the classification of the utterance *Okay* as an acceptance or acknowledgment. It represents a confusion matrix describing the number of occasions that coders used pairs of labels for a given turn. This table shows that the two coders favored the use of *accept* strongly over *acknowledge*. They correctly state that this skew in the distribution of categories increases the expected chance agreement, thus lowering the overall agreement in chance-corrected tests. The reason for this is that since one category is more popular than others, the likelihood of coders' agreeing by chance by choosing this category increases. We therefore require a comparable increase in observed agreement to accommodate this.

Di Eugenio and Glass (2004) perceive this as an "unpleasant behavior" of chance-corrected tests, one that prevents us from concluding that the example given in Table 1 shows satisfactory levels of agreement. Instead they use percentage agreement to arrive at this conclusion. By examining the data, it is clear that this conclusion would be false.

In Table 1, the coders agree 90 out of 100 times, but all agreements occur when both coders choose *accept*. There is not a single case in which they agree on *Okay*'s being used as an acknowledgment. The only conclusion one may justifiably draw is that the coders cannot distinguish the use of *Okay* as an acceptance from its use as an acknowledgment. Rather than being an unpleasant behavior, accounting for prevalence in the data is an

**Table 1**
Prevalence in coding.

| | | Coder 2 | | |
|---|---|---|---|---|
| Coder 1 | | Accept | Ack | |
| | Accept | 90 | 5 | 95 |
| | Acknowledge | 5 | 0 | 5 |
| | | 95 | 5 | 100 |

important part of accurately reporting the level of agreement. This helps us to avoid arriving at incorrect conclusions such as believing that the data shown in Table 1 suggest reliable coding.

## 2.2 Chance-Corrected Agreement: Unequal Coder Category Distribution

The second class of agreement measure recommended in Di Eugenio and Glass (2004) is that of **chance-corrected tests** that do not assume an equal distribution of categories between coders. Chance-corrected tests compute agreement according to the ratio of observed (dis)agreement to that which we could expect by chance, estimated from the data. The measures differ in the way in which this expected (dis)agreement is estimated. Those that do not assume an equal distribution between coders calculate expected (dis)agreement based on the individual distribution of each coder.

The concern that in discourse and dialogue coding, coders will differ in the frequency with which they apply labels leads Di Eugenio and Glass to conclude that Cohen's (1960) kappa is the best chance-corrected test to apply. To clarify, by unequal distribution of categories, we do not refer to the disparity in the frequency with which categories occur (e.g., verbs are more common than pronouns) but rather to the difference in proclivity between coders (e.g., coder A is more likely to label something a noun than coder B).

Cohen's kappa calculates expected chance agreement, based on the individual coders' distributions, in a manner similar to association measures, such as chi–square. This means that its results are dependent on the preferences of the individual coders taking part in the tests. This violates the condition set out at the beginning of this section whereby agreement must be a function of the coding process, with coders being viewed as interchangeable. The purpose of assessing the reliability of coding schemes is not to judge the performance of the small number of individuals participating in the trial, but rather to predict the performance of the schemes in general. The proposal that in most discourse and dialogue studies, the assumption of equal distribution between coders does not hold is, in fact, an argument *against* the use of Cohen's kappa. Assessing the agreement between coders and accounting for their idiosyncratic proclivity toward or against certain labels tells us little about how the coding scheme will perform when applied by others. The solution is not to apply a test that panders to individual differences, but rather to increase the number of coders so that the influence of any individual on the final result becomes less pronounced.[1]

Another reason provided for using Cohen's kappa is that its sensitivity to bias (differences in coders' category distribution) can be exploited to improve coding schemes. However, there is no need to calculate kappa in order to observe bias, since it will be evident in a contingency table of the data in question. Even if it were necessary to compute kappa for this purpose, however, this would not justify its use as a reliability test.

## 2.3 Chance-Corrected Agreement: Assumed Equal Coder Category Distribution

The remaining class of agreement measure assumes an equal distribution of categories for all coders. Once we have accepted that this assumption is necessary in order to

[1] When there is a single *correct* label that should be used, such as part-of-speech tags used to describe the syntactic function of a word or group of words, then training coders may mitigate coder preference.

predict the performance of the scheme in general, there appears to be no objection to using this type of statistical test for assessing agreement in discourse and dialogue work. Tests that fall into this class include Siegel and Castellan's (1988) extension of Scott's (1955) pi, confusingly called kappa, and Krippendorff's (2004a) alpha. Both of these measures calculate expected (dis)agreement based on the frequency with which each category is used, estimated from the overall usage by the coders.

Kappa is more frequently described in statistics textbooks and more commonly implemented in statistical software. In circumstances in which mechanisms other than nominal labels are used to annotate data, alpha has the benefit of being able to deal with different degrees of disagreement between pairs of interval, ordinal, and ratio values, among others.

Di Eugenio and Glass (2004) conclude with the proposal that these three forms of agreement measure collectively provide better means with which to judge agreement than any individual test. We would argue, to the contrary, that applying three different metrics to measure the same property suggests a lack of confidence in any of them. Percentage agreement and Cohen's kappa do not provide an insight into a scheme's reliability, so reporting their results is potentially misleading.

## 3. Inferring Reliability

To reiterate, when testing reliability we are assessing whether the data that a scheme generates can be relied on. This may be inferred from the level of agreement between coders applying the scheme. In section 1 we described two properties of reliable data that are important to establish in discourse and dialogue analysis. In this section we explain how the gap between agreement and reliability may be bridged.

When inferring reliability from agreement, a common error is to believe that there are a number of thresholds against which agreement scores can be measured in order to gauge whether or not a coding scheme produces reliable data. Most commonly this is Krippendorff's decision criterion, in which scores greater than 0.8 are considered satisfactory and scores greater than 0.667 allow tentative conclusions to be drawn (Krippendorff 2004a). The prevalent use of this criterion despite repeated advice that it is unlikely to be suitable for all studies (Carletta 1996; Di Eugenio and Glass 2004; Krippendorff 2004a) is probably due to a desire for a simple system that can be easily applied to a scheme. Unfortunately, because of the diversity of both the phenomena being coded and the applications of the results, it is impossible to prescribe a scale against which all coding schemes can be judged.

Instead we provide discussion and some recommendations, all founded on the premise that reliability must "correlate with the conditions under which one is willing to rely on imperfect data" (Krippendorff 2004b, page 6). A common concern regarding the application of standards from other fields, such as the one described above, to discourse and dialogue research is that the subjectivity of the phenomena being coded may mean that we never obtain the necessary agreement levels. In this context, **subjectivity** describes the absence of an obvious mapping for each unit of analysis onto categories that describe the phenomenon in question. However, the fact that we consider these subjective phenomena worthy of study shows that we are, in fact, "willing to rely on imperfect data," which is fine as long as we recognize the limitations of a scheme that delivers less-than-ideal levels of reliability and use the resulting annotated corpora accordingly.

In order to discuss the acceptable levels of agreement for discourse and dialogue coding, let us consider two popular uses of coded data: to train systems to perform some automated task and to study the relationship between the coded phenomena and some other feature of the data.

### 3.1 Reliability and Training for Automatic Annotation

Considering the effort involved in manually annotating linguistic data, it is unsurprising that attempts are often made to train a system to perform such annotation automatically (Mast et al. 1996; Wrede and Shriberg 2003). The reliability of manually annotated data is clearly a concern when they are used to train a system. If the level of agreement for the annotation scheme is low, then the system is going to replicate the inconsistent behavior of human annotators. Any deviant behavior by the system resulting in less than 100% accuracy in comparison with the manual annotation will compound the problem, possibly leading to meaningless data. Worse still, if a system is to learn how to annotate from manually annotated data, it will do so based on the patterns observed in those data. If the manual annotation is not reliable, then those patterns may be nonexistent or misleading.

Returning to our original premise, we would suggest that if a coding scheme is to be used to generate data from which a system will learn to perform similar coding, then we should be "unwilling to rely on imperfect data."

### 3.2 Reliability and Corpus Analysis

Manually annotated corpora can also be used to infer a relationship between the phenomena in question and some other facet of the data. When performing this sort of analysis, we may be more willing to work with imperfect data and therefore accept lower levels of agreement. However, the conclusions that are gleaned from the analysis must be tempered according to the level of agreement achieved. For example, when it is suggested that a correlation exists between the occurrence of one phenomenon and that of another, less agreement observed in the sample annotation requires stronger evidence of the correlation in order for the conclusion to be valid.

To summarize, there are no magic thresholds that, once crossed, entitle us to claim that a coding scheme is reliable. One must decide for oneself, based on the intended use of a scheme, whether the observed level of agreement is sufficient and conduct one's analysis accordingly.

## 4. Conclusion

The application of agreement statistics has done much to improve the scientific rigor of discourse and dialogue research. However, unless we understand what we are attempting to prove and which tests are appropriate, the results of evaluation can be unsatisfactory or, worse still, misleading. In this article we have encouraged researchers to clarify their reasons for assessing agreement and have suggested that in many cases the most suitable test for this purpose is one that corrects for expected agreement, based on an assumed equal distribution between coders.

### Acknowledgments

The authors acknowledge the help of Klaus Krippendorff for patiently aiding our understanding of reliability and University of Manchester for funding Richard Craggs.

### References

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 43(6):37–46.

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual*. Technical Report (Draft 13), University of Colorado.

Krippendorff, Klaus. 2004a. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Sage, Beverly Hills, CA.

Krippendorff, Klaus. 2004b. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–437.

Mast, Marion, Heinrich Niemann, Elmar Noth, and Ernst Gunter Schukat-Talamazzini. 1996. Automatic classification of dialog acts with semantic classification trees and polygrams. In *Learning for Natural Language Processing*, edited by Stefan Wermter, Ellen Riloff, and Gabriele Scheler. Springer, pages 217–229.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19:127–141.

Siegel, Sidney and John N. Castellan, Jr. 1988. *Nonparametric Statistics*. 2nd ed. McGraw-Hill.

Wrede, Britta and Elizabeth Shriberg. 2003. Spotting “hot spots” in meetings: Human judgments and prosodic cues. In *Proceedings of EUROSPEECH*, Geneva.