Applying Reliability Metrics to Co-Reference Annotation

Rebecca J. Passonneau
Columbia University
Department of Computer Science
Technical Report CUCS-017-97
June 5, 1997

Abstract

Studies of the contextual and linguistic factors that constrain discourse phenomena such as reference are coming to depend increasingly on annotated language corpora. In preparing the corpora, it is important to evaluate the reliability of the annotation, but methods for doing so have not been readily available. In this report, I present a method for computing reliability of coreference annotation. First I review a method for applying the information retrieval metrics of recall and precision to coreference annotation proposed by Marc Vilain and his collaborators. I show how this method makes it possible to construct contingency tables for computing Cohen's κ, a familiar reliability metric. By comparing recall and precision to reliability on the same data sets, I also show that recall and precision can be misleadingly high. Because κ factors out chance agreement among coders, it is a preferable measure for developing annotated corpora where no pre-existing target annotation exists.

1 Two Reliability Metrics

Two equivalent metrics for quantifying inter-rater reliability between pairs of coders are Cohen's κ coefficient of agreement (1960) and Krippendorff's α (1980). The formulas for each are shown in (1) and (2).

$$\kappa = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \tag{1}$$

$$\alpha = 1 - \frac{p_{D_O}}{p_{D_E}} \tag{2}$$

$$1 = p_{A_O} + p_{D_O} \tag{3}$$

$$1 = p_{A_E} + p_{D_E} \tag{4}$$

$$\frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} = 1 - \frac{p_{D_O}}{p_{D_E}} \tag{5}$$

$$\frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} = 1 - \frac{(1 - p_{A_O})}{(1 - p_{A_E})} \tag{6}$$

$$\frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} = \frac{(1 - p_{A_E})}{(1 - p_{A_E})} - \frac{(1 - p_{A_O})}{(1 - p_{A_E})} \tag{7}$$

$$\frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} = \frac{p_{A_O} - p_{A_E}}{(1 - p_{A_E})} \tag{8}$$

Briefly, Cohen's κ is cast in terms of the amount of *agreement* between coders that exceeds chance expectations. The numerator of the ratio in (1) is the proportion of observed agreements (p_{A_O}) less the proportion expected to agree by chance (p_{A_E}); the denominator is the total proportion (100%) less the the proportion expected to agree by chance. Conversely, Krippendorff's α is cast in terms of the extent to which the observed *disagreements* between coders is below chance expectation; it is the total probability less the ratio of observed disagreements to expected disagreements. The observed probability of agreement and disagreement must sum to one, as must the expected probability of agreement and disagreement ((3) and (4)). By substitution, it can be shown that κ equals α ((5) - (8)).

The reliability measures depend crucially on a hypothesis of chance expectation. In (Cohen, 1960) and (Krippendorff, 1980), chance expectation is derived from the marginals of

Judge X	Judge Y A	B	
A	47	14	*61*
B	10	29	*39*
	57	*43*	*100*

$$\alpha \quad = \kappa \quad = .50 \tag{9}$$

Table 1: A 2-by-2 coincidence matrix

a coincidence matrix classifying the response categories of one coder by the response categories of another coder. Table 1 illustrates a simple 2-by-2 coincidence matrix. A coincidence matrix classifies a set of data in a way that shows, for a given set of classification categories (e.g., A versus B), how the data is cross-classified. Every data point must go in one and only one cell of the table to indicate how the data classified by one coding (row categories) is cross-classified by the other coding (column categories). The diagonal from upper left to lower right in Table 1 represents the responses of judge X that *coincide* with judge Y's; cells off the diagonal represent classification disagreements.

The marginals in Table 1 show that 61% of judge X's responses are in category A compared with 57% of Y's. Where .61 is taken to be the likelihood that X responds in category A, and .57 the likelihood that Y responds in category A, then $.61 \times .57$ of the time X and Y should agree that the same data point is classified in category A, assuming nothing more than chance correspondence between X and Y's responses. Adding the result of the corresponding likelihood of agreement on response B yields $p_{A_E} = 52\%$. The expected proportion of disagreement is similarly computed. By chance, X should respond A where Y responds B 26% of the time ($.61 \times .43$). The difference between these expected values and the observed agreements ($.47 + .29$) results in a reliability value of .50, as shown in (9) of Table 1.

Whenever the responses of two subjects can be cast in the form of a coincidence matrix, the reliability metrics illustrated above can be applied. Here I present a proposal for applying reliability to coreference annotation, based on the insights in (Vilain et al., 1995).

2 Evaluating Coreference Annotations

Co-reference annotation is annotation of language data to indicate when distinct expressions have been used to corefer. Evaluating the reliability of such data is important for several reasons. First, any annotation task is subject to unintended errors arising from lack of attention on the part of the annotator. The likelihood of such errors depends in part on ergonomic factors such as what kinds of aids are provided for recording and checking annotations, and how much time the annotator has to perform the task. In addition, no matter how precise a language user might be, language interpretation is subjective. A given expression can be referentially ambiguous or vague. Referential indeterminacy can even be intentional on the part of the speaker or writer. When annotations of the same data are collected from two or more coders, then in principle, the reliability of the data (or of the individual coders) can be quantified.

Two language samples are presented in Figure 1 that typify two quite different types of discourse. Sample 1 illustrates journalistic text, and is taken from the Brown Corpus (Francis and Kucera, 1982). Sample 2, illustrating spoken dialogue, is from the University of Rochester's Trains 91 corpus (Gross et al., 1993). Two samples are

Sample 1: Journalistic Text

Committee approval of [Gov. Price Daniel's *abandoned property* act]$_1$ seemed certain Thursday despite the adamant protests of Texas bankers. [Daniel]$_2$ personally led the fight for [the measure]$_1$, which [he]$_2$ had watered down considerably since its rejection by two previous Legislatures, in a public hearing before the House Committee on Revenue and Taxation. Under committee rules, [it]$_1$ went automatically to a subcommittee for one week. But questions with which [committee members]$_3$ taunted bankers appearing as witnesses left little doubt that [they]$_3$ will recommend passage of [it]$_1$.

Sample 2: Problem-Solving Dialogue

M: okay we need to ship a boxcar of oranges to Bath by 8 AM today S: okay M: umm okay so I guess uh I would suggest that we use [engine E1]$_1$ uh and have [it]$_1$ pick up [a boxcar]$_2$ at ah Dansville how long'll [it]$_1$ take S: uh that'll take 3 hours to get to Dansville and get [the boxcar]$_2$ M: uh okay and then how long to go on to .. Corning with [the boxcar]$_2$ coupled to uh [E1]$_1$ S: another hour M: ok so that's okay and then uh if we loaded [the oranges]$_3$ at ah Corning and sent ah [E1]$_1$ on to Bath with [the oranges]$_3$ S: we'd get there at 7

Co-reference Annotations

Token	String	CA
A	Gov. Price Daniel's ... act	1
B	Daniel	2
C	the measure	1
D	he	2
E	it	1
F	comittee members	3
G	they	3
H	it	1

Token	String	CA_1	CA_2
A	engine E1	1	1
B	it	1	1
C	a boxcar	2	2
D	it	1	2
E	the boxcar	2	2
F	the boxcar	2	2
G	E1	1	1
H	the oranges	3	3
I	E1	1	1
J	the oranges	3	3

Figure 1: Co-reference annotation of two language samples

shown to illustrate that despite major differences of language variety, the task of coreference annotation is essentially the same for both types of data. Both samples have been annotated to indicate certain expressions that have been interpreted to corefer (how or why these particular expressions were selected is immaterial to the present discussion). Relevant phrases have been bracketed. Bracketed phrases that have been annotated with the same numeric subscript represent expressions that, in the annotator's judgement, were used to corefer. For sample 1, eight expressions (A-H) were annotated as referring to one of three distinct referents. The coding of co-referential expressions is shown under column CA (Coreference Annotation). For sample 2, ten expressions (A-J) were annotated as referring to one of three distinct referents, whose indices are listed under the column headed CA_1. An alternate coding is shown in column CA_2. The remainder of the discussion will focus on sample 2.

How can a comparison of the two annotations of sample 2, CA_1 and CA_2, be quantified? The key observations used in (Vilain et al., 1995) are that the sets of expressions that corefer constitute equivalence classes, and that in two annotations, a given expression is either assigned to the same equivalence class or not. I first present how (Vilain et al., 1995) compute precision and recall by comparing equivalence classes across a pair of annotations. Then I show how a revision of their approach can be converted to reliability measures, under certain important constraints.

The first annotation for Sample 2 places five tokens into one equivalence class referring to the engine ({A, B, D, G, I}), and three tokens into a class referring to the boxcar ({C, E, F}). This contrasts with the alternate annotation, where the same eight tokens are in two equivalence classes, but where D is placed with C: {A, B, G, I}, {C, D, E, F}. To apply recall and precision, we must assume that one of the annotations is *correct.* In general, a recall error involves failure to identify members of a target set; a precision error involves inclusion of additional elements besides those in the target set. Vilain et al. (1995) observe that intuitively, a comparison of two sets {A, B, D, G, I} from CA_1 and {A, B, G, I} from CA_2, where the first set is the target, involves only a recall error. The CA_2 set does not include any additional elements, but it fails to include D. In contrast, the comparison of {C, E, F} as the target with {C, D, E, F} involves a precision error and no recall errors. In practice, the method given in (Vilain et al., 1995) does not compare elements of corresponding sets, but compares how many links are needed to connect the elements within corresonding sets.

To compute recall, (Vilain et al., 1995) start by creating a partition of a given target set from the corresponding response sets. This addresses the question of how many equivalence classes in the response set must be examined in order to reconstruct the target set. The relevant partition of {A, B, D, G, I} is thus into the two sets {A, B, G, I}, {D}. If the target set is conceived of as five nodes in a spanning tree (e.g., A–B–D–G–I), then the target "tree" can be constructed from the response by adding one link: a link from D to any node A, B, G or I. In general, the missing information for recall is quantified in terms of the number of links missing from the response partition. The number of links in a target equivalence class C is the cardinality of that class less 1: $|C| - 1$. The number of links missing from the partition of C relative to the response ($p(C)$) is the cardinality of the partition less 1: $|p(C)| - 1$. The recall for a given equivalence class is thus the ratio of the target links less the missing links to the target links:

$$\begin{aligned} Recall_C &= \frac{(|C| - 1) - (|p(C)| - 1)}{(|C| - 1)} \\ &= \frac{|C| - |p(C)|}{(|C| - 1)} \end{aligned} \tag{10}$$

When an equivalence class C_i in the target has an exact correspondence to one in the response, the cardinality of the partition $p(C_i)$ is 1, the numerator and denominator in (10) are the same, and recall is perfect. Recall for a complete annotation is expressed in terms of all the equivalences classes C_i in the target annotation, by summing the recall errors (numerator) and summing the target links (denominator):

$$Recall = \frac{\sum_i |C_i| - |p(C_i)|}{\sum_i (|C_i| - 1)} \tag{11}$$

$$Recall_{CA_1,CA_2} \quad \frac{(5-2)+(3-1)+(2-1)}{(5-1)+(3-1)+(2-1)} \tag{12}$$

Taking CA_1 as the target, formula (11) gives a recall for CA_2 of .86, as shown in (12).

Computation of precision in (Vilain et al., 1995) is the converse of the computation of recall. To illustrate, precision will be computed for the target set {C, E, F}. Precision is imperfect because the response set has an additional member: {C, D, E, F}. Where the *response* set is R, a partition of the response set relative to the target sets ($p(R)$) gives the two sets {C, E, F} and {D}. Precision of the *target* set C is then the ratio of the difference between the cardinality of the corresponding response set R and the cardinality of its partition p(R) to the cardinality of the response set R less 1:

$$Precision_C = \frac{|R| - |p(R)|}{(|R| - 1)} \tag{13}$$

M: okay we need to ship a boxcar of oranges to Bath by 8 AM today S: okay M: umm okay so I guess uh I would suggest that we use [engine E1]$_1$ uh and have [it]$_1$ pick up [a boxcar]$_2$ at ah Dansville how long'll [it]$_3$ take S: uh [that]$_3$'ll take 3 hours to get to Dansville and get [the boxcar]$_2$ M: uh okay and then how long to go on to .. Corning with [the boxcar]$_2$ coupled to uh [E1]$_1$ S: another hour M: ok so that's okay and then uh if we loaded [the oranges]$_4$ at ah Corning and sent ah [E1]$_1$ on to Bath with [the oranges]$_4$ S: we'd get there at 7

Token	String	CA_1	CA_3	CA_1 Equivalence classes	CA_3 Equvalence classes
A	engine E1	1	1	{A, B, D, G, I}	{A, B, G, I}
B	it	1	1	{C, E, F}	{C, E, F}
C	a boxcar	2	2	{H, J}	{D, D'}
D	it	1	3		{H, J}
D'	that	4	3		
E	the boxcar	2	2		
F	the boxcar	2	2		
G	E1	1	1		
H	the oranges	3	4		
I	E1	1	1		
J	the oranges	3	4		

Figure 2: Alternate co-reference annotation of sample 2

$$Precision = \frac{\sum_i |R_i| - |p(R_i)|}{\sum_i (|R_i| - 1)} \quad (14)$$

$$Precision_{CA_1,CA_2} = \frac{(4-1)+(4-2)+(2-1)}{(4-1)+(4-1)+(2-1)} \quad (15)$$

Precision for the equivalence class {C, E, F} is then $\frac{4-2}{4-1}$, or .33. Precision of the entire coding CA_2 relative to CA_1 is .86, as shown in (15).

2.1 Problems

A perhaps more realistic alternate coding for sample 2 is shown in Figure 2. The token identified in Figures 1-2 as D was coded as coreferential with the expression *engine E1* (token A) in annotation CA_1. In annotation CA_3 shown in Figure 2, this token is interpreted to refer to the process of getting engine E1 to pick up a boxcar at Dansville, and is annotated as coreferential with a token of the demonstrative pronoun *that*—shown here as token D'. D' was not originally included in CA_1, but is given here an arbitrary index of 4 in coding CA_1 to indicate lack of coreference with any other expression. I will use a comparison of codings CA_1 and CA_3 to illustrate how the approach taken in (Vilain et al., 1995) presents certain problems for computing reliability, and for evaluating the type of annotation employed in (Passonneau and Litman, 1997).

Both of the problems discussed here pertain to the manner in which recall and precision is applied to data, rather than to the actual computation of recall and precision. The first problem is that (Vilain et al., 1995) do not constrain the sets of referring expressions that are being compared to have the same cardinality. The second is that they apply their method only to referring expressions that corefer with at least one other expression. My proposed solution requires that two annotations have the same cardinality of referring expressions. It also permits an annotator to interpret an expression as having no coreferential expressions, as in D' for coding CA_1 (Figure 2). As I show below, these two moves make it possible to retain the basic insight from (Vilain et al., 1995), to compute reliability, and to apply the method to a broader range of annotation approaches, including the annotation style presented in (Passonneau, 1997).

The fundamental problem in comparing codings CA_1 and CA_3 is that the two data sets are incommensurate. Coding CA_1 originally placed ten expressions into equivalence classes, while coding CA_3 does so for eleven expressions. This prevents creation of a contingency table, and is thus an obstacle to applying reliability measures (cf. section 1).

The approach in (Vilain et al., 1995) does not require two codings to be commensurate in part because the annotators' task, as described in (Hirschman, 1996), has two parts: to identify the expressions to be coded, or *markables*, and to place markables into equivalence classes based on the coreference relation. As I argue in (Passonneau, 1997), there are several disadvantages to this approach. Identifying markables is a conceptually distinct task, can be partly automated with easily accessible and relatively simple tools, such as part-of-speech taggers, and is a language specific task. In contrast, coreference is difficult to automate (particularly in a sufficiently general way to apply across corpora), and is language independent. I take the evaluation of how markables are identified to be a separate problem. My goal is then to evaluate the inter-rater reliability of co-reference annotations, assuming that each rater is given the same set of markables to annotate.

Another serious drawback, of particular concern to investigators in the natural language generation community, is that the approach taken in (Vilain et al., 1995) fails to identify referential expressions comprising a singleton equivalence class. Instead, such expressions are omitted from consideration. However, it is of as much concern to determine the conditions under which a referent is mentioned only once, as to determine those under which it is re-mentioned. If two coders place the same expression in a class by itself, indicating lack of any coreferential expressions, note that recall and precision will both be zero. While at first this may seem counter-intuitive, it is entirely reasonable. First, what is being evaluated is the ability of distinct coders to find the same coreference links. In the case of comparing a singleton set to an identical singleton set, there are no coreference links to find. But note that no mismatching links have been identified.

	Coding CA_1		
Coding CA_3	+Link	-Link	
+Link	a	b	a+b
-Link	c	d	c+d
	a+c	b+d	a+b+c+d

$$Recall = \frac{a}{a+c} = \frac{\sum_i |C_i| - |p(C_i)|}{\sum_i (|C_i| - 1)} \quad (16)$$

$$Precision = \frac{a}{a+b} = \frac{\sum_i |C'_i| - |p(C'_i)|}{\sum_i (|C'_i| - 1)} \quad (17)$$

Figure 3: Schematic representation of a 2-by-2 coincidence matrix

Consider the result of imposing the requirement that two coreference codings must partition the same set of expressions into equivalence classes of coreference. If we assume that coding CA_1 represents an annotator's judgement that token D' is in a singleton set, then we can create a contingency table of the two codings. The table total represents the total number of possible coreference links. In the case of codings CA_1 and CA_3, the table total is the cardinality of the set of tokens less 1, which is ten. To compute reliability, we need the four quantities $a - d$ given in each cell of the table shown in Figure 3 (cf. Table 1). Of all possible coreference links, some will be identified by both coders. This is quantity a in Figure 3. Some will be identified by neither coder: quantity d in

Figure 3. Thus a and d represent the two types of agreement between coders: agreement on coreference links, and agreement on their absence. In contrast, quantities b and c represent disagreements: the first coder finds links that the second coder does not, or vice versa.

Recall and precision are defined as illustrated in (16) and (17) of Figure 3 (Rijsbergen, 1979). Recall represents the ratio of links found in both the target and some test set, hence is the ratio of a to $a + c$. By setting this ratio equal to (10), the ratio proposed in (Vilain et al., 1995), we can begin to identify the individual quantities a through d. Precision represents the proportion of links found in some test set that are also in the target, hence is the ratio of a to $a + b$. As shown in (17), this ratio can be equated to (14). Given the table total and the two equalities (16) and (17), the four quantities a through d can be computed.

Recall that quantity a is the coreference links agreed on by CA_1 and CA_3. By (16) and (17), it is the sum of the differences of the cardinality of each equivalence class in CA_1 less the cardinality of its partition by CA_3. Equivalently, a is the sum of the differences of the cardinality of each equivalence class in CA_3 less the cardinality of its partition by corresponding equivalence classes in CA_1:

$$
\begin{aligned}
a = & \quad (5-2)+(3-1)+(1-1)+(2-1) \\
a = & \quad (4-1)+(3-1)+(2-2)+(2-1) \\
a = & \quad 6
\end{aligned}
$$

Cell value b represents the coreference links identified in CA_3 but not in CA_1. It is the sum of the number of links for each equivalence class in CA_3 ($\sum_i |C_i| - 1$) less the coreference links found by both:

$$
\begin{aligned}
b = & \quad ((4-1)+(3-1)+(2-1)+(2-1))-6 \\
b = & \quad 1
\end{aligned}
$$

Conversely, cell value c represents the coreference links identified in CA_1 but not in CA_3. It is the sum of the number of links for each equivalence class in CA_1 less the coreference links found by both:

$$
\begin{aligned}
c = & \quad ((5-1)+(3-1)+(1-1)+(2-1))-6 \\
c = & \quad 1
\end{aligned}
$$

It remains to calculate d, the possible links that neither coder identifies. We know the total possible coreference links: $a+b+c+d = 10$. And we know the values of a, b and c (a=6; b=c=1), thus d = 2. Another way to compute a and d is to compute the full partition of the equivalence classes in both codings (p(CA)), giving all links found in both codings:

p(CA) = {A, B, G, I}, {C, E, F}, {D}, {D'}, {H, J}

Note that the value of a (links agreed on by both coders) is the sum of the differences of the cardinality of each set in the partition p(CA) less 1:

$$
\begin{aligned}
a = & \quad (4-1)+(3-1)+(1-1)+(1-1)+(2-1) \\
a = & \quad 6
\end{aligned}
$$

Then take the intersection of either CA_1 or CA_3 with p(CA). The value of d is the cardinality of either intersection less 1:

$$CA_1 \cap p(CA) \quad = \quad \{C, E, F\}, \{D'\}, \{H, J\}$$

Coding CA_3	Coding CA_1 +Link	-Link	
+Link	6	1	7
-Link	1	2	3
	7	3	10

$$Recall = \frac{6}{7} = 85.7\% \quad (18)$$
$$Precision = \frac{6}{7} = 85.7\% \quad (19)$$

Table 2: Coincidence matrix for CA_1 by CA_3

$$\begin{aligned} CA_3 \cap p(CA) &= \{A,B,G,I\},\{C,E,F\},\{H,J\} \\ d &= |CA_1 \cap p(CA)| - 1 \\ d &= |CA_3 \cap p(CA)| - 1 \\ d &= 2 \end{aligned}$$

The contingency table for comparing CA_1 and CA_3 using the cell values we have just computed is given in Table 2.

2.2 Conversion to Reliability

Now that we see how to construct a contingency table for coreference annotation, it is straightforward to compute reliability. Given that recall and precision are both just over 85%, one might interpret the similarity of the coding as being moderately good. However, as shown in (20)-(28), reliability is poor. The interpretation of the κ value of .52 is that reliability is about halfway between completely random behavior (*kappa* = 0) and perfect reliability (near 1).[1]

$$\kappa = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \quad (20)$$
$$p_{A_O} = .6 + .2 \quad (21)$$
$$p_{A_E} = (.7 \times .7) + (.3 \times .3) \quad (22)$$

[1]A negative *kappa* value represents positive unreliability, as opposed to random correspondence. See (Cohen, 1960) for a discussion of the upper and lower limits of κ assuming p_{A_E} is derived from marginals of a coincidence matrix. See (Krippendorff, 1980) for other methods of computing p_{A_E}, and for applying reliability to continuous variables, etc.

Coding R_2	Coding R_1 +Link	-Link	
+Link	166	19	185
-Link	13	44	57
	179	63	242

$$Recall = .90\% \quad (29)$$
$$Precision = .93\% \quad (30)$$

$$\kappa = .65 \quad (31)$$

Table 3: Coincidence matrix for R_1 by R_2

$$\kappa = \frac{(.6 + .2) - ((.7 \times .7) + (.3 \times .3))}{1 - ((.7 \times .7) + (.3 \times .3))} \quad (23)$$
$$\alpha = 1 - \frac{p_{D_O}}{p_{D_E}} \quad (24)$$
$$p_{D_O} = .1 + .1 \quad (25)$$
$$p_{D_E} = (.7 \times .3) + (.7 \times .3) \quad (26)$$
$$\alpha = 1 - \frac{(.1 + .1)}{((.7 \times .3) + (.7 \times .3))} \quad (27)$$
$$\kappa = \alpha = .52 \quad (28)$$

Table 3 compares the κ reliability score with recall and precision for an actual coding of a spoken narrative from (Chafe, 1980). One coding represents the consensus coding of coreference arrived at by the two investigators in the study reported in (Passonneau and Litman, 1997). The other coding was performed by a student with no linguistics background but some training in coreference annotation. As illustrated, the recall and precision scores are both apparently good (90% or above), but the κ score is only .65. This demonstrates concretely that because recall and precision do not factor out chance agreement, they can be misleading. In contrast, as discussed in section 1, κ quantifies the proportion of agreements among two coders that are above chance. In Table 3, both coders agree on 166 out of 242 coreference links (upper left cell). Because of the relatively high value of this cell, both recall and precision will be high (cf. Figure 3). But in addition, because the proportion of coreference links is very high for both R_1 (179/242) and R_2 (185/242), the chance of agreement on coreference links (or their absence) is also relatively high. Factoring out this chance agreement results in poor reliability.

Table 4 compares the κ scores with recall and precision for the same coder's annotations of ten narratives from (Chafe, 1980) against the codings used in (Passonneau and Litman, 1997). Narrative one, with a κ of .85 compared with recall and precision of .96, illustrates the general trend that the κ scores are good, but not as high as one might assume given the generally high recall and precision. The last line of the table gives the standard deviation (σ) for each metric. Note that the standard deviation of the reliability measures is over 3 times that for recall and precision. A log kept by the coder of questions that arose during annotation suggests that the variation in reliability reflects differences in the coherence of the narratives, and the types of referential phenomena that occur, rather than inconsistency in the coder's behavior. For example, in this log the coder reported greatest difficulty with narratives 9 (α=.75) and 12 (α=.74), and used the phrases "*I am confused, I don't understand what he is talking about*" to describe particular coding problems. In contrast, the coder described narrative 16 (α=.93) as "*pretty easy to code.*"

Narr.	κ	Recall	Precision
1	.85	.96	.96
2	.65	.90	.93
3	.72	.93	.94
4	.89	.94	.98
5	.89	.95	.99
6	.83	.94	.97
8	.84	.91	.96
9	.75	.88	.96
11	.79	.92	.95
12	.74	.90	.92
15	.80	.93	.93
16	.93	.97	.98
17	.86	.95	.96
18	.84	.93	.96
19	.85	.96	.93
σ	.07	.02	.02

Table 4: Comparing Inter-rater Reliability of Coreference Annotations with Recall and Precision

3 Summary

A 2-by-2 coincidence matrix can be used to compute information retrieval metrics, or to compute reliability. Building on this observation, I have shown how the method in Vilain et al. (1995) for computing recall and precision for coreference annotation can be used to construct a coincidence matrix, and therefore to compute reliability. Each type of metric has its own uses. If a target or correct annotation has been established, it may be appropriate to evaluate recall and precision of a new coding against the target. However, in developing new annotated corpora with no pre-exising answer key, so to speak, it is important to evaluate the reliability of individual coders and of the datasets they produce. The data presented in the preceding section (Tables 3-4) demonstrate that one should not infer from high recall and precision of one annotation against another that either annotation is reliable, in the sense of reliability discussed in (Cohen, 1960) and (Krippendorff, 1980). Reliability measures should be used to identify reliable annotators and annotations. By merging the best data from mutually reliable codings, a more correct coding can be derived for a new corpus. Reliability scores can be used to determine whether a coder is trainable (improvements over time), and when the training can be terminated (no further improvement).

Poor reliability can be an indicator of omissions or flaws in a coding scheme. In addition, reliability metrics can help the researcher identify data that is consistently not agreed upon among multiple coders. This might occur within a single discourse for particular kinds of coreference phenomena. Or it might occur for an entire discourse as compared with other discourses, e.g., if the discourse in question is unclear, vague, or otherwise non-optimal for coreference interpretation.

Acknowledgements

Thanks to Pamela Jordan, Diane Litman, and Marilyn Walker for helpful comments on this report.

References

Wallace L. Chafe. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ablex Publishing Corporation, Norwood, NJ.

Jacob Cohen. 1960. A coeffiecient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

W. Francis and H. Kucera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston, MA.

Derek Gross, James Allen, and David R. Traum. 1993. The Trains 91 dialogues. Technical Report 92-1, University of Rochester, Rochester, NY.

Lynette Hirschman. 1996. Coreference specification for MUC (MUCCS). Version 4.0, Oct. 29, 1996.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Beverly Hills, CA.

Rebecca J. Passonneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23.1:103–139. Special Issue on Empirical Studies in Discourse Interpretation and Generation.

Rebecca J. Passonneau. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Technical report, Columbia University.

C. J. Van Rijsbergen. 1979. *Information Retrieval.* Butterworths, London.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, pages 45–52, San Francisco. Morgan Kaufmann.