

# Squibs and Discussions

## The Kappa Statistic: A Second Look

Barbara Di Eugenio\*  
University of Illinois at Chicago

Michael Glass†  
Valparaiso University

*In recent years, the kappa coefficient of agreement has become the de facto standard for evaluating intercoder agreement for tagging tasks. In this squib, we highlight issues that affect  $\kappa$  and that the community has largely neglected. First, we discuss the assumptions underlying different computations of the expected agreement component of  $\kappa$ . Second, we discuss how prevalence and bias affect the  $\kappa$  measure.*

In the last few years, coded corpora have acquired an increasing importance in every aspect of human-language technology. Tagging for many phenomena, such as dialogue acts (Carletta et al. 1997; Di Eugenio et al. 2000), requires coders to make subtle distinctions among categories. The objectivity of these decisions can be assessed by evaluating the reliability of the tagging, namely, whether the coders reach a satisfying level of agreement when they perform the same coding task. Currently, the de facto standard for assessing intercoder agreement is the  $\kappa$  coefficient, which factors out expected agreement (Cohen 1960; Krippendorff 1980).  $\kappa$  had long been used in content analysis and medicine (e.g., in psychiatry to assess how well students' diagnoses on a set of test cases agree with expert answers) (Grove et al. 1981). Carletta (1996) deserves the credit for bringing  $\kappa$  to the attention of computational linguists.

$\kappa$  is computed as  $\frac{P(A) - P(E)}{1 - P(E)}$ , where  $P(A)$  is the observed agreement among the coders, and  $P(E)$  is the expected agreement, that is,  $P(E)$  represents the probability that the coders agree by chance. The values of  $\kappa$  are constrained to the interval  $[-1, 1]$ . A  $\kappa$  value of one means perfect agreement, a  $\kappa$  value of zero means that agreement is equal to chance, and a  $\kappa$  value of negative one means "perfect" disagreement.

This squib addresses two issues that have been neglected in the computational linguistics literature. First, there are two main ways of computing  $P(E)$ , the expected agreement, according to whether the distribution of proportions over the categories is taken to be equal for the coders (Scott 1955; Fleiss 1971; Krippendorff 1980; Siegel and Castellan 1988) or not (Cohen 1960). Clearly, the two approaches reflect different conceptualizations of the problem. We believe the distinction between the two is often glossed over because in practice the two computations of  $P(E)$  produce very similar outcomes in most cases, especially for the highest values of  $\kappa$ . However, first, we will show that they can indeed result in different values of  $\kappa$ , that we will call  $\kappa_{Co}$  (Cohen 1960) and  $\kappa_{S\&C}$  (Siegel and Castellan 1988). These different values can lead to contradictory conclusions on intercoder agreement. Moreover, the assumption of

---

\* Computer Science, 1120 SEO (M/C 152), 851 South Morgan Street, Chicago, IL 60607. E-mail: bdiugen@uic.edu.

† Mathematics and Computer Science, 116 Gellerson Hall, Valparaiso, IN 46383. E-mail: michael.glass@valpo.edu.

equal distributions over the categories masks the exact source of disagreement among the coders. Thus, such an assumption is detrimental if such systematic disagreements are to be used to improve the coding scheme (Wiebe, Bruce, and O'Hara 1999).

Second,  $\kappa$  is affected by skewed distributions of categories (the **prevalence problem**) and by the degree to which the coders disagree (the **bias problem**). That is, for a fixed  $P(A)$ , the values of  $\kappa$  vary substantially in the presence of prevalence, bias, or both.

We will conclude by suggesting that  $\kappa_{Co}$  is a better choice than  $\kappa_{S\&C}$  in those studies in which the assumption of equal distributions underlying  $\kappa_{S\&C}$  does not hold: the vast majority, if not all, of discourse- and dialogue-tagging efforts. However, as  $\kappa_{Co}$  suffers from the bias problem but  $\kappa_{S\&C}$  does not,  $\kappa_{S\&C}$  should be reported too, as well as a third measure that corrects for prevalence, as suggested in Byrt, Bishop, and Carlin (1993).

### 1. The Computation of $P(E)$

$P(E)$  is the probability of agreement among coders due to chance. The literature describes two different methods for estimating a probability distribution for random assignment of categories. In the first, each coder has a personal distribution, based on that coder's distribution of categories (Cohen 1960). In the second, there is one distribution for all coders, derived from the total proportions of categories assigned by all coders (Scott 1955; Fleiss 1971; Krippendorff 1980; Siegel and Castellan 1988).<sup>1</sup>

We now illustrate the computation of  $P(E)$  according to these two methods. We will then show that the resulting  $\kappa_{Co}$  and  $\kappa_{S\&C}$  may straddle one of the significant thresholds used to assess the raw  $\kappa$  values.

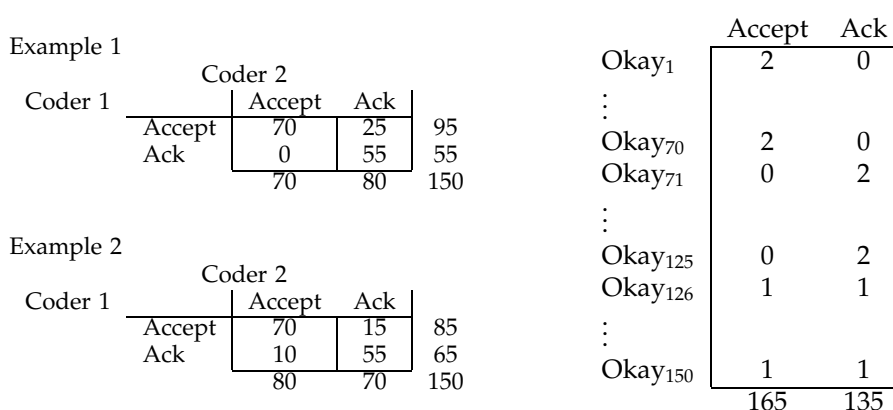
The assumptions underlying these two methods are made tangible in the way the data are visualized, in a **contingency table** for Cohen, and in what we will call an **agreement table** for the others. Consider the following situation. Two coders<sup>2</sup> code 150 occurrences of *Okay* and assign to them one of the two labels *Accept* or *Acknowledgement* (Allen and Core 1997). The two coders label 70 occurrences as *Accept*, and another 55 as *Ack*. They disagree on 25 occurrences, which one coder labels as *Ack*, and the other as *Accept*. In Figure 1, this example is encoded by the top contingency table on the left (labeled *Example 1*) and the agreement table on the right. The contingency table directly mirrors our description. The agreement table is an  $N \times m$  matrix, where  $N$  is the number of items in the data set and  $m$  is the number of labels that can be assigned to each object; in our example,  $N = 150$  and  $m = 2$ . Each entry  $n_{ij}$  is the number of codings of label  $j$  to item  $i$ . The agreement table in Figure 1 shows that occurrences 1 through 70 have been labeled as *Accept* by both coders, 71 through 125 as *Ack* by both coders, and 126 to 150 differ in their labels.

<sup>1</sup> To be precise, Krippendorff uses a computation very similar to Siegel and Castellan's to produce a statistic called alpha. Krippendorff computes  $P(E)$  (called  $1 - D_e$  in his terminology) with a sampling-without-replacement methodology. The computations of  $P(E)$  and of  $1 - D_e$  show that the difference is negligible:

$$P(E) = \sum_j \left( \frac{\sum_i n_{ij}}{Nk} \right)^2 \quad (\text{Siegel and Castellan})$$

$$1 - D_e = \sum_j \left( \frac{\sum_i n_{ij}}{Nk} \right) \left( \frac{[\sum_i n_{ij}] - 1}{Nk - 1} \right) \quad (\text{Krippendorff})$$

<sup>2</sup> Both  $\kappa_{S\&C}$  (Scott 1955) and  $\kappa_{Co}$  (Cohen 1960) were originally devised for two coders. Each has been extended to more than two coders, for example, respectively Fleiss (1971) and Bartko and Carpenter (1976). Thus, without loss of generality, our examples involve two coders.



**Figure 1** Cohen’s contingency tables (left) and Siegel and Castellan’s agreement table (right).

Agreement tables lose information. When the coders disagree, we cannot reconstruct which coder picked which category. Consider Example 2 in Figure 1. The two coders still disagree on 25 occurrences of *Okay*. However, one coder now labels 10 of those as *Accept* and the remaining 15 as *Ack*, whereas the other labels the same 10 as *Ack* and the same 15 as *Accept*. The agreement table does not change, but the contingency table does.

Turning now to computing  $P(E)$ , Figure 2 shows, for Example 1, Cohen’s computation of  $P(E)$  on the left, and Siegel and Castellan’s computation on the right. We include the computations of  $\kappa_{Co}$  and  $\kappa_{S\&C}$  as the last step. For both Cohen and Siegel and Castellan,  $P(A) = 125/150 = 0.8333$ . The observed agreement  $P(A)$  is computed as the proportion of items the coders agree on to the total number of items;  $N$  is the number of items, and  $k$  the number of coders ( $N = 150$  and  $k = 2$  in our example). Both  $\kappa_{Co}$  and  $\kappa_{S\&C}$  are highly significant at the  $p = 0.5 * 10^{-5}$  level (significance is computed for  $\kappa_{Co}$  and  $\kappa_{S\&C}$  according to the formulas in Cohen [1960] and Siegel and Castellan [1988], respectively).

The difference between  $\kappa_{Co}$  and  $\kappa_{S\&C}$  in Figure 2 is just under 1%, however, the results of the two  $\kappa$  computations straddle the value 0.67, which for better or worse has been adopted as a cutoff in computational linguistics. This cutoff is based on the assessment of  $\kappa$  values in Krippendorff (1980), which discounts  $\kappa < 0.67$  and allows tentative conclusions when  $0.67 \leq \kappa < 0.8$  and definite conclusions when  $\kappa \geq 0.8$ . Krippendorff’s scale has been adopted without question, even though Krippendorff himself considers it only a plausible standard that has emerged from his and his colleagues’ work. In fact, Carletta et al. (1997) use words of caution against adopting Krippendorff’s suggestion as a standard; the first author has also raised the issue of how to assess  $\kappa$  values in Di Eugenio (2000).

If Krippendorff’s scale is supposed to be our standard, the example just worked out shows that the different computations of  $P(E)$  do affect the assessment of inter-coder agreement. If less-strict scales are adopted, the discrepancies between the two  $\kappa$  computations play a larger role, as they have a larger effect on smaller values of  $\kappa$ . For example, Rietveld and van Hout (1993) consider  $0.20 < \kappa \leq 0.40$  as indicating fair agreement, and  $0.40 < \kappa \leq 0.60$  as indicating moderate agreement. Suppose that two coders are coding 100 occurrences of *Okay*. The two coders label 40 occurrences as *Accept* and 25 as *Ack*. The remaining 35 are labeled as *Ack* by one coder and as *Accept* by the other (as in Example 6 in Figure 4);  $\kappa_{Co} = 0.418$ , but  $\kappa_{S\&C} = 0.27$ . These two values are really at odds.

Assumption of different distributions among coders (Cohen)

**Step 1.** For each category  $j$ , compute the overall proportion  $p_{j,l}$  of items assigned to  $j$  by each coder  $l$ . In a contingency table, each row and column total divided by  $N$  corresponds to one such proportion for the corresponding coder.

$$p_{Accept,1} = 95/150, p_{Ack,1} = 55/150, \\ p_{Accept,2} = 70/150, p_{Ack,2} = 80/150$$

**Step 2.** For a given item, the likelihood of both coders' independently agreeing on category  $j$  by chance, is  $p_{j,1} * p_{j,2}$ .

$$p_{Accept,1} * p_{Accept,2} = 95/150 * 70/150 = 0.2956 \\ p_{Ack,1} * p_{Ack,2} = 55/150 * 80/150 = 0.1956$$

**Step 3.**  $P(E)$ , the likelihood of coders' accidentally assigning the same category to a given item, is  $\sum_j p_{j,1} * p_{j,2} = 0.2956 + 0.1956 = 0.4912$

**Step 4.**  
 $\kappa_{Co} = (0.8333 - 0.4912)/(1 - 0.4912) = .3421/.5088 = 0.6724$

Assumption of equal distributions among coders (Siegel and Castellan)

**Step 1.** For each category  $j$ , compute  $p_j$ , the overall proportion of items assigned to  $j$ . In an agreement table, the column totals give the total counts for each category  $j$ , hence:

$$p_j = \frac{1}{Nk} \times \sum_i n_{ij}$$

$$p_{Accept} = 165/300 = 0.55, p_{Ack} = 135/300 = 0.45$$

**Step 2.** For a given item, the likelihood of both coders' independently agreeing on category  $j$  by chance is  $p_j^2$ .

$$p_{Accept}^2 = 0.3025 \\ p_{Ack}^2 = 0.2025$$

**Step 3.**  $P(E)$ , the likelihood of coders' accidentally assigning the same category to a given item, is  $\sum_j p_j^2 = 0.3025 + 0.2025 = 0.5050$

**Step 4.**  
 $\kappa_{S\&C} = (0.8333 - 0.5050)/(1 - 0.5050) = .3283/.4950 = 0.6632$

**Figure 2**

The computation of  $P(E)$  and  $\kappa$  according to Cohen (left) and to Siegel and Castellan (right).

## 2. Unpleasant Behaviors of Kappa: Prevalence and Bias

In the computational linguistics literature,  $\kappa$  has been used mostly to validate coding schemes: Namely, a “good” value of  $\kappa$  means that the coders agree on the categories and therefore that those categories are “real.” We noted previously that assessing what constitutes a “good” value for  $\kappa$  is problematic in itself and that different scales have been proposed. The problem is compounded by the following obvious effect on  $\kappa$  values: If  $P(A)$  is kept constant, varying values for  $P(E)$  yield varying values of  $\kappa$ . What can affect  $P(E)$  even if  $P(A)$  is constant are prevalence and bias.

The prevalence problem arises because skewing the distribution of categories in the data increases  $P(E)$ . The minimum value  $P(E) = 1/m$  occurs when the labels are equally distributed among the  $m$  categories (see Example 4 in Figure 3). The maximum value  $P(E) = 1$  occurs when the labels are all concentrated in a single category. But for a given value of  $P(A)$ , the larger the value of  $P(E)$ , the lower the value of  $\kappa$ .

Example 3 and Example 4 in Figure 3 show two coders agreeing on 90 out of 100 occurrences of *Okay*, that is,  $P(A) = 0.9$ . However,  $\kappa$  ranges from  $-0.048$  to  $0.80$ , and from not significant to significant (the values of  $\kappa_{S\&C}$  for Examples 3 and 4 are the same as the values of  $\kappa_{Co}$ ).<sup>3</sup> The differences in  $\kappa$  are due to the difference in the relative prevalence of the two categories *Accept* and *Ack*. In Example 3, the distribution is skewed, as there are 190 *Accepts* but only 10 *Acks* across the two coders; in Example 4, the distribution is even, as there are 100 *Accepts* and 100 *Acks*, respectively. These results do not depend on the size of the sample; that is, they are not due to the fact

<sup>3</sup> We are not including agreement tables for the sake of brevity.

Example 3

		Coder 2		
		Accept	Ack	
Coder 1	Accept	90	5	95
	Ack	5	0	5
		95	5	100

$$P(A) = 0.90, P(E) = 0.905$$

$$\kappa_{Co} = \kappa_{S\&C} = -0.048, p = 1$$

Example 4

		Coder 2		
		Accept	Ack	
Coder 1	Accept	45	5	50
	Ack	5	45	50
		50	50	100

$$P(A) = 0.90, P(E) = 0.5$$

$$\kappa_{Co} = \kappa_{S\&C} = 0.80, p = 0.5 * 10^{-5}$$

**Figure 3**

Contingency tables illustrating the prevalence effect on  $\kappa$ .

Example 5

		Coder 2		
		Accept	Ack	
Coder 1	Accept	40	15	55
	Ack	20	25	45
		60	40	100

$$P(A) = 0.65, P(E) = 0.52$$

$$\kappa_{Co} = 0.27, p = 0.005$$

Example 6

		Coder 2		
		Accept	Ack	
Coder 1	Accept	40	35	75
	Ack	0	25	25
		40	60	100

$$P(A) = 0.65, P(E) = 0.45$$

$$\kappa_{Co} = 0.418, p = 0.5 * 10^{-5}$$

**Figure 4**

Contingency tables illustrating the bias effect on  $\kappa_{Co}$ .

Example 3 and Example 4 are small. As the computations of  $P(A)$  and  $P(E)$  are based on proportions, the same distributions of categories in a much larger sample, say, 10,000 items, will result in exactly the same  $\kappa$  values. Although this behavior follows squarely from  $\kappa$ 's definition, it is at odds with using  $\kappa$  to assess a coding scheme. From both Example 3 and Example 4 we would like to conclude that the two coders are in substantial agreement, independent of the skewed prevalence of *Accept* with respect to *Ack* in Example 3. The role of prevalence in assessing  $\kappa$  has been subject to heated discussion in the medical literature (Grove et al. 1981; Berry 1992; Goldman 1992).

The bias problem occurs in  $\kappa_{Co}$  but not  $\kappa_{S\&C}$ . For  $\kappa_{Co}$ ,  $P(E)$  is computed from each coder's individual probabilities. Thus, the less two coders agree in their overall behavior, the fewer chance agreements are expected. But for a given value of  $P(A)$ , decreasing  $P(E)$  will increase  $\kappa_{Co}$ , leading to the paradox that  $\kappa_{Co}$  increases as the coders become less similar, that is, as the marginal totals diverge in the contingency table. Consider two coders coding the usual 100 occurrences of *Okay*, according to the two tables in Figure 4. In Example 5, the proportions of each category are very similar among coders, at 55 versus 60 *Accept*, and 45 versus 40 *Ack*. However, in Example 6 coder 1 favors *Accept* much more than coder 2 (75 versus 40 occurrences) and conversely chooses *Ack* much less frequently (25 versus 60 occurrences). In both cases,  $P(A)$  is 0.65 and  $\kappa_{S\&C}$  is stable at 0.27, but  $\kappa_{Co}$  goes from 0.27 to 0.418. Our initial example in Figure 1 is also affected by bias. The distribution in Example 1 yielded  $\kappa_{Co} = 0.6724$  but  $\kappa_{S\&C} = 0.6632$ . If the bias decreases as in Example 2,  $\kappa_{Co}$  becomes 0.6632, the same as  $\kappa_{S\&C}$ .

### 3. Discussion

The issue that remains open is which computation of  $\kappa$  to choose. Siegel and Castellan's  $\kappa_{S\&C}$  is not affected by bias, whereas Cohen's  $\kappa_{Co}$  is. However, it is

questionable whether the assumption of equal distributions underlying  $\kappa_{S\&C}$  is appropriate for coding in discourse and dialogue work. In fact, it appears to us that it holds in few if any of the published discourse- or dialogue-tagging efforts for which  $\kappa$  has been computed. It is, for example, appropriate in situations in which item  $i$  may be tagged by different coders than item  $j$  (Fleiss 1971). However,  $\kappa$  assessments for discourse and dialogue tagging are most often performed on the same portion of the data, which has been annotated by each of a small number of annotators (between two and four). In fact, in many cases the analysis of systematic disagreements among annotators on the same portion of the data (i.e., of bias) can be used to improve the coding scheme (Wiebe, Bruce, and O'Hara 1999).

To use  $\kappa_{Co}$  but to guard against bias, Cicchetti and Feinstein (1990) suggest that  $\kappa_{Co}$  be supplemented, for each coding category, by two measures of agreement, positive and negative, between the coders. This means a total of  $2m$  additional measures, which we believe are too many to gain a general insight into the meaning of the specific  $\kappa_{Co}$  value. Alternatively, Byrt, Bishop, and Carlin (1993) suggest that intercoder reliability be reported as three numbers:  $\kappa_{Co}$  and two adjustments of  $\kappa_{Co}$ , one with bias removed, the other with prevalence removed. The value of  $\kappa_{Co}$  adjusted for bias turns out to be  $\dots \kappa_{S\&C}$ . Adjusted for prevalence,  $\kappa_{Co}$  yields a measure that is equal to  $2P(A) - 1$ . The results for Example 1 should then be reported as  $\kappa_{Co} = 0.6724$ ,  $\kappa_{S\&C} = 0.6632$ ,  $2P(A) - 1 = 0.6666$ ; those for Example 6 as  $\kappa_{Co} = 0.418$ ,  $\kappa_{S\&C} = 0.27$ , and  $2P(A) - 1 = 0.3$ . For both Examples 3 and 4,  $2P(A) - 1 = 0.8$ . Collectively, these three numbers appear to provide a means of better judging the meaning of  $\kappa$  values. Reporting both  $\kappa$  and  $2P(A) - 1$  may seem contradictory, as  $2P(A) - 1$  does *not* correct for expected agreement. However, when the distribution of categories is skewed, this highlights the effect of prevalence. Reporting both  $\kappa_{Co}$  and  $\kappa_{S\&C}$  does not invalidate our previous discussion, as we believe  $\kappa_{Co}$  is more appropriate for discourse- and dialogue-tagging in the majority of cases, especially when exploiting bias to improve coding (Wiebe, Bruce, and O'Hara 1999).

### Acknowledgments

This work is supported by grant N00014-00-1-0640 from the Office of Naval Research. Thanks to Janet Cahn and to the anonymous reviewers for comments on earlier drafts.

### References

- Allen, James and Mark Core. 1997. DAMSL: Dialog act markup in several layers; Coding scheme developed by the participants at two discourse tagging workshops, University of Pennsylvania, March 1996, and Schloß Dagstuhl, February 1997. Draft.
- Bartko, John J. and William T. Carpenter. 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163(5):307–317.
- Berry, Charles C. 1992. The  $\kappa$  statistic [letter to the editor]. *Journal of the American Medical Association*, 268(18):2513–2514.
- Byrt, Ted, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429.
- Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.
- Cicchetti, Domenic V. and Alvan R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.
- Di Eugenio, Barbara. 2000. On the usage of Kappa to evaluate agreement on coding tasks. In *LREC2000: Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 441–444, Athens.
- Di Eugenio, Barbara, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 2000. The agreement process: An

- empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human Computer Studies*, 53(6):1017–1076.
- Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Goldman, Ronald L. 1992. The  $\kappa$  statistic [letter to the editor (in reply)]. *Journal of the American Medical Association*, 268(18):2513–2514.
- Grove, William M., Nancy C. Andreasen, Patricia McDonald-Scott, Martin B. Keller, and Robert W. Shapiro. 1981. Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, 38:408–413.
- Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.
- Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin.
- Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19:127–141.
- Siegel, Sidney and N. John Castellan, Jr. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw Hill, Boston.
- Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *ACL99: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, MD.