Survey Article

Inter-Coder Agreement for Computational Linguistics

Ron Artstein*
University of Essex

Massimo Poesio**
University of Essex/Università di Trento

This article is a survey of methods for measuring agreement among corpus annotators. It exposes the mathematics and underlying assumptions of agreement coefficients, covering Krippendorff's alpha as well as Scott's pi and Cohen's kappa; discusses the use of coefficients in several annotation tasks; and argues that weighted, alpha-like coefficients, traditionally less used than kappa-like measures in computational linguistics, may be more appropriate for many corpus annotation tasks—but that their use makes the interpretation of the value of the coefficient even harder.

1. Introduction and Motivations

Since the mid 1990s, increasing effort has gone into putting semantics and discourse research on the same empirical footing as other areas of computational linguistics (CL). This soon led to worries about the subjectivity of the judgments required to create annotated resources, much greater for semantics and pragmatics than for the aspects of language interpretation of concern in the creation of early resources such as the Brown corpus (Francis and Kucera 1982), the British National Corpus (Leech, Garside, and Bryant 1994), or the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993). Problems with early proposals for assessing coders' agreement on discourse segmentation tasks (such as Passonneau and Litman 1993) led Carletta (1996) to suggest the adoption of the K coefficient of agreement, a variant of Cohen's κ (Cohen 1960), as this had already been used for similar purposes in content analysis for a long time.[1] Carletta's proposals

* Now at the Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina Del Rey, CA 90292.

** At the University of Essex: Department of Computing and Electronic Systems, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. E-mail: poesio@essex.ac.uk. At the University of Trento: CIMeC, Università degli Studi di Trento, Palazzo Fedrigotti, Corso Bettini, 31, 38068 Rovereto (TN), Italy. E-mail: massimo.poesio@unitn.it.

1 The literature is full of terminological inconsistencies. Carletta calls the coefficient of agreement she argues for "kappa," referring to Krippendorff (1980) and Siegel and Castellan (1988), and using Siegel and Castellan's terminology and definitions. However, Siegel and Castellan's statistic, which they call K, is actually Fleiss's generalization to more than two coders of Scott's π, not of the original Cohen's κ; to confuse matters further, Siegel and Castellan use the Greek letter κ to indicate the parameter which is estimated by K. In what follows, we use κ to indicate Cohen's original coefficient and its generalization to more than two coders, and K for the coefficient discussed by Siegel and Castellan.

Submission received: 26 August 2005; revised submission received: 21 December 2007; accepted for publication: 28 January 2008.

were enormously influential, and K quickly became the de facto standard for measuring agreement in computational linguistics not only in work on discourse (Carletta et al. 1997; Core and Allen 1997; Hearst 1997; Poesio and Vieira 1998; Di Eugenio 2000; Stolcke et al. 2000; Carlson, Marcu, and Okurowski 2003) but also for other annotation tasks (e.g., Véronis 1998; Bruce and Wiebe 1998; Stevenson and Gaizauskas 2000; Craggs and McGee Wood 2004; Mieskes and Strube 2006). During this period, however, a number of questions have also been raised about K and similar coefficients—some already in Carletta's own work (Carletta et al. 1997)—ranging from simple questions about the way the coefficient is computed (e.g., whether it is really applicable when more than two coders are used), to debates about which levels of agreement can be considered 'acceptable' (Di Eugenio 2000; Craggs and McGee Wood 2005), to the realization that K is not appropriate for all types of agreement (Poesio and Vieira 1998; Marcu, Romera, and Amorrortu 1999; Di Eugenio 2000; Stevenson and Gaizauskas 2000). Di Eugenio raised the issue of the effect of **skewed distributions** on the value of K and pointed out that the original κ developed by Cohen is based on very different assumptions about coder bias from the K of Siegel and Castellan (1988), which is typically used in CL. This issue of annotator bias was further debated in Di Eugenio and Glass (2004) and Craggs and McGee Wood (2005). Di Eugenio and Glass pointed out that the choice of calculating chance agreement by using individual coder marginals (κ) or pooled distributions (K) can lead to reliability values falling on different sides of the accepted 0.67 threshold, and recommended reporting both values. Craggs and McGee Wood argued, following Krippendorff (2004a,b), that measures like Cohen's κ are inappropriate for measuring agreement. Finally, Passonneau has been advocating the use of Krippendorff's α (Krippendorff 1980, 2004a) for coding tasks in CL which do not involve nominal and disjoint categories, including anaphoric annotation, wordsense tagging, and summarization (Passonneau 2004, 2006; Nenkova and Passonneau 2004; Passonneau, Habash, and Rambow 2006).

Now that more than ten years have passed since Carletta's original presentation at the workshop on Empirical Methods in Discourse, it is time to reconsider the use of coefficients of agreement in CL in a systematic way. In this article, a survey of coefficients of agreement and their use in CL, we have three main goals. First, we discuss in some detail the mathematics and underlying assumptions of the coefficients used or mentioned in the CL and content analysis literatures. Second, we also cover in some detail Krippendorff's α, often mentioned but never really discussed in detail in previous CL literature other than in the papers by Passonneau just mentioned. Third, we review the past ten years of experience with coefficients of agreement in CL, reconsidering the issues that have been raised also from a mathematical perspective.[2]

2. Coefficients of Agreement

2.1 Agreement, Reliability, and Validity

We begin with a quick recap of the goals of agreement studies, inspired by Krippendorff (2004a, Section 11.1). Researchers who wish to use hand-coded data—that is, data in which **items** are labeled with **categories**, whether to support an empirical claim or to develop and test a computational model—need to show that such data are **reliable**.

2 Only part of our material could fit in this article. An extended version of the survey is available from http://cswww.essex.ac.uk/Research/nle/arrau/.

The fundamental assumption behind the methodologies discussed in this article is that data are reliable if coders can be shown to **agree** on the categories assigned to units to an extent determined by the purposes of the study (Krippendorff 2004a; Craggs and McGee Wood 2005). If different coders produce consistently similar results, then we can infer that they have internalized a similar understanding of the annotation guidelines, and we can expect them to perform consistently under this understanding.

Reliability is thus a prerequisite for demonstrating the **validity** of the coding scheme—that is, to show that the coding scheme captures the "truth" of the phenomenon being studied, in case this matters: If the annotators are not consistent then either some of them are wrong or else the annotation scheme is inappropriate for the data. (Just as in real life, the fact that witnesses to an event disagree with each other makes it difficult for third parties to know what actually happened.) However, it is important to keep in mind that achieving good agreement cannot ensure validity: Two observers of the same event may well share the same prejudice while still being objectively wrong.

2.2 A Common Notation

It is useful to think of a reliability study as involving a set of **items** (markables), a set of **categories**, and a set of **coders** (annotators) who assign to each item a unique category label. The discussions of reliability in the literature often use different notations to express these concepts. We introduce a uniform notation, which we hope will make the relations between the different coefficients of agreement clearer.

- The set of **items** is $\{ i \mid i \in I \}$ and is of cardinality **i**.
- The set of **categories** is $\{ k \mid k \in K \}$ and is of cardinality **k**.
- The set of **coders** is $\{ c \mid c \in C \}$ and is of cardinality **c**.

Confusion also arises from the use of the letter P, which is used in the literature with at least three distinct interpretations, namely "proportion," "percent," and "probability." We will use the following notation uniformly throughout the article.

- A_o is observed agreement and D_o is observed disagreement.
- A_e and D_e are expected agreement and expected disagreement, respectively. The relevant coefficient will be indicated with a superscript when an ambiguity may arise (for example, A_e^{π} is the expected agreement used for calculating π, and A_e^{κ} is the expected agreement used for calculating κ).
- $P(\cdot)$ is reserved for the probability of a variable, and $\hat{P}(\cdot)$ is an estimate of such probability from observed data.

Finally, we use **n** with a subscript to indicate the number of judgments of a given type:

- $\mathbf{n}_{ik}$ is the number of coders who assigned item i to category k;
- $\mathbf{n}_{ck}$ is the number of items assigned by coder c to category k;
- $\mathbf{n}_{k}$ is the total number of items assigned by all coders to category k.

2.3 Agreement Without Chance Correction

The simplest measure of agreement between two coders is **percentage of agreement** or **observed agreement**, defined for example by Scott (1955, page 323) as "the percentage of judgments on which the two analysts agree when coding the same data independently." This is the number of items on which the coders agree divided by the total number of items. More precisely, and looking ahead to the following discussion, observed agreement is the arithmetic mean of the **agreement value** agr_i for all items $i \in I$, defined as follows:

$$\text{agr}_i = \begin{cases} 1 \text{ if the two coders assign } i \text{ to the same category} \\ 0 \text{ if the two coders assign } i \text{ to different categories} \end{cases}$$

Observed agreement over the values agr_i for all items $i \in I$ is then:

$$A_o = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i$$

For example, let us assume a very simple annotation scheme for dialogue acts in information-seeking dialogues which makes a binary distinction between the categories `statement` and `info-request`, as in the DAMSL dialogue act scheme (Allen and Core 1997). Two coders classify 100 utterances according to this scheme as shown in Table 1. Percentage agreement for this data set is obtained by summing up the cells on the diagonal and dividing by the total number of items: $A_o = (20 + 50)/100 = 0.7$.

Observed agreement enters in the computation of all the measures of agreement we consider, but on its own it does not yield values that can be compared across studies, because some agreement is due to chance, and the amount of chance agreement is affected by two factors that vary from one study to the other. First of all, as Scott (1955, page 322) points out, "[percentage agreement] is biased in favor of dimensions with a small number of categories." In other words, given two coding schemes for the same phenomenon, the one with fewer categories will result in higher percentage agreement just by chance. If two coders randomly classify utterances in a uniform manner using the scheme of Table 1, we would expect an equal number of items to fall in each of the four cells in the table, and therefore pure chance will cause the coders to agree on half of the items (the two cells on the diagonal: $\frac{1}{4} + \frac{1}{4}$). But suppose we want to refine the simple binary coding scheme by introducing a new category, `check`, as in the MapTask coding scheme (Carletta et al. 1997). If two coders randomly classify utterances in a uniform manner using the three categories in the second scheme, they would only agree on a third of the items ($\frac{1}{9} + \frac{1}{9} + \frac{1}{9}$).

Table 1
A simple example of agreement on dialogue act tagging.

		CODER A		
		STAT	IREQ	TOTAL
CODER B	STAT	20	20	40
	IREQ	10	50	60
	TOTAL	30	70	100

The second reason percentage agreement cannot be trusted is that it does not correct for the distribution of items among categories: We expect a higher percentage agreement when one category is much more common than the other. This problem, already raised by Hsu and Field (2003, page 207) among others, can be illustrated using the following example (Di Eugenio and Glass 2004, example 3, pages 98–99). Suppose 95% of utterances in a particular domain are `statement`, and only 5% are `info-request`. We would then expect by chance that $0.95 \times 0.95 = 0.9025$ of the utterances would be classified as `statement` by both coders, and $0.05 \times 0.05 = 0.0025$ as `info-request`, so the coders would agree on 90.5% of the utterances. Under such circumstances, a seemingly high observed agreement of 90% is actually worse than expected by chance.

The conclusion reached in the literature is that in order to get figures that are comparable across studies, observed agreement has to be adjusted for chance agreement. These are the measures we will review in the remainder of this article. We will not look at the variants of percentage agreement used in CL work on discourse before the introduction of kappa, such as percentage agreement with an expert and percentage agreement with the majority; see Carletta (1996) for discussion and criticism.[3]

2.4 Chance-Corrected Coefficients for Measuring Agreement between Two Coders

All of the coefficients of agreement discussed in this article correct for chance on the basis of the same idea. First we find how much agreement is expected by chance: Let us call this value A_e. The value $1 - A_e$ will then measure how much agreement over and above chance is attainable; the value $A_o - A_e$ will tell us how much agreement beyond chance was actually found. The ratio between $A_o - A_e$ and $1 - A_e$ will then tell us which proportion of the possible agreement beyond chance was actually observed. This idea is expressed by the following formula.

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

The three best-known coefficients, S (Bennett, Alpert, and Goldstein 1954), π (Scott 1955), and κ (Cohen 1960), and their generalizations, all use this formula; whereas Krippendorff's α is based on a related formula expressed in terms of disagreement (see Section 2.6). All three coefficients therefore yield values of agreement between $-A_e/1 - A_e$ (no observed agreement) and 1 (observed agreement = 1), with the value 0 signifying chance agreement (observed agreement = expected agreement). Note also that whenever agreement is less than perfect ($A_o < 1$), chance-corrected agreement will be strictly lower than observed agreement, because some amount of agreement is always expected by chance.

Observed agreement A_o is easy to compute, and is the same for all three coefficients—the proportion of items on which the two coders agree. But the notion of chance agreement, or the probability that two coders will classify an arbitrary item as belonging to the same category by chance, requires a model of what would happen if coders' behavior was only by chance. All three coefficients assume *independence* of the two coders—that is, that the chance of c_1 and c_2 agreeing on any given category k

3 The extended version of the article also includes a discussion of why χ^2 and correlation coefficients are not appropriate for this task.

Table 2
The value of different coefficients applied to the data from Table 1.

Coefficient	Expected agreement	Chance-corrected agreement
S	$2 \times (\frac{1}{2})^2 = 0.5$	$(0.7 - 0.5)/(1 - 0.5) = 0.4$
π	$0.35^2 + 0.65^2 = 0.545$	$(0.7 - 0.545)/(1 - 0.545) \approx 0.341$
κ	$0.3 \times 0.4 + 0.6 \times 0.7 = 0.54$	$(0.7 - 0.54)/(1 - 0.54) \approx 0.348$

Observed agreement for all the coefficients is 0.7.

is the product of the chance of each of them assigning an item to that category: $\mathrm{P}(k|c_1) \cdot \mathrm{P}(k|c_2)$.[4] Expected agreement is then the probability of c_1 and c_2 agreeing on any category, that is, the sum of the product over all categories:

$$\mathrm{A}_\mathrm{e}^S = \mathrm{A}_\mathrm{e}^\pi = \mathrm{A}_\mathrm{e}^\kappa = \sum_{k \in K} \mathrm{P}(k|c_1) \cdot \mathrm{P}(k|c_2)$$

The difference between S, π, and κ lies in the assumptions leading to the calculation of $\mathrm{P}(k|c_i)$, the chance that coder c_i will assign an arbitrary item to category k (Zwick 1988; Hsu and Field 2003).

- S: This coefficient is based on the assumption that if coders were operating by chance alone, we would get a uniform distribution: That is, for any two coders c_m, c_n and any two categories k_j, k_l, $\mathrm{P}(k_j|c_m) = \mathrm{P}(k_l|c_n)$.
- π: If coders were operating by chance alone, we would get the same distribution for each coder: For any two coders c_m, c_n and any category k, $\mathrm{P}(k|c_m) = \mathrm{P}(k|c_n)$.
- κ: If coders were operating by chance alone, we would get a separate distribution for each coder.

Additionally, the lack of independent prior knowledge of the distribution of items among categories means that the distribution of categories (for π) and the priors for the individual coders (for κ) have to be estimated from the observed data. Table 2 demonstrates the effect of the different chance models on the coefficient values. The remainder of this section explains how the three coefficients are calculated when the reliability data come from two coders; we will discuss a variety of proposed generalizations starting in Section 2.5.

*2.4.1 All Categories Are Equally Likely: **S**.* The simplest way of discounting for chance is the one adopted to compute the coefficient S (Bennett, Alpert, and Goldstein 1954), also known in the literature as C, κ_n, G, and RE (see Zwick 1988; Hsu and Field 2003). As noted previously, the computation of S is based on an interpretation of chance as a random choice of category from a uniform distribution—that is, all categories are equally likely. If coders classify the items into **k** categories, then the chance $\mathrm{P}(k|c_i)$ of

4 The independence assumption has been the subject of much criticism, for example by John S. Uebersax. `http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm`.

any coder assigning an item to category k under the uniformity assumption is $\frac{1}{\mathbf{k}}$; hence the total agreement expected by chance is

$$A_e^S = \sum_{k \in K} \frac{1}{\mathbf{k}} \cdot \frac{1}{\mathbf{k}} = \mathbf{k} \cdot \left(\frac{1}{\mathbf{k}}\right)^2 = \frac{1}{\mathbf{k}}$$

The calculation of the value of S for the figures in Table 1 is shown in Table 2.

The coefficient S is problematic in many respects. The value of the coefficient can be artificially increased simply by adding spurious categories which the coders would never use (Scott 1955, pages 322–323). In the case of CL, for example, S would reward designing extremely fine-grained tagsets, provided that most tags are never actually encountered in real data. Additional limitations are noted by Hsu and Field (2003). It has been argued that uniformity is the best model for a chance distribution of items among categories if we have no independent prior knowledge of the distribution (Brennan and Prediger 1981). However, a lack of prior knowledge does not mean that the distribution cannot be estimated post hoc, and this is what the other coefficients do.

2.4.2 A Single Distribution: π. All of the other methods for discounting chance agreement we discuss in this article attempt to overcome the limitations of S's strong uniformity assumption using an idea first proposed by Scott (1955): Use the actual behavior of the coders to estimate the prior distribution of the categories. As noted earlier, Scott based his characterization of π on the assumption that random assignment of categories to items, by any coder, is governed by the distribution of items among categories in the actual world. The best estimate of this distribution is $\hat{P}(k)$, the observed proportion of items assigned to category k by both coders.

$$P(k|c_1) = P(k|c_2) = \hat{P}(k)$$

$\hat{P}(k)$, the observed proportion of items assigned to category k by both coders, is the total number of assignments to k by both coders $\mathbf{n}_k$, divided by the overall number of assignments, which for the two-coder case is twice the number of items $\mathbf{i}$:

$$\hat{P}(k) = \frac{\mathbf{n}_k}{2\mathbf{i}}$$

Given the assumption that coders act independently, expected agreement is computed as follows.

$$A_e^\pi = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = \sum_{k \in K} \left(\frac{\mathbf{n}_k}{2\mathbf{i}}\right)^2 = \frac{1}{4\mathbf{i}^2} \sum_{k \in K} \mathbf{n}_k^2$$

It is easy to show that for any set of coding data, $A_e^\pi \geq A_e^S$ and therefore $\pi \leq S$, with the limiting case (equality) obtaining when the observed distribution of items among categories is uniform.

2.4.3 Individual Coder Distributions: κ. The method proposed by Cohen (1960) to calculate expected agreement A_e in his κ coefficient assumes that random assignment of categories to items is governed by prior distributions that are unique to each coder, and which reflect individual **annotator bias**. An individual coder's prior distribution is

estimated by looking at her actual distribution: $P(k|c_i)$, the probability that coder c_i will classify an arbitrary item into category k, is estimated by using $\hat{P}(k|c_i)$, the proportion of items actually assigned by coder c_i to category k; this is the number of assignments to k by c_i, $\mathbf{n}_{c_i k}$, divided by the number of items $\mathbf{i}$.

$$P(k|c_i) = \hat{P}(k|c_i) = \frac{\mathbf{n}_{c_i k}}{\mathbf{i}}$$

As in the case of S and π, the probability that the two coders c_1 and c_2 assign an item to a particular category $k \in K$ is the joint probability of each coder making this assignment independently. For κ this joint probability is $\hat{P}(k|c_1) \cdot \hat{P}(k|c_2)$; expected agreement is then the sum of this joint probability over all the categories $k \in K$.

$$A_e^\kappa = \sum_{k \in K} \hat{P}(k|c_1) \cdot \hat{P}(k|c_2) = \sum_{k \in K} \frac{\mathbf{n}_{c_1 k}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 k}}{\mathbf{i}} = \frac{1}{\mathbf{i}^2} \sum_{k \in K} \mathbf{n}_{c_1 k} \mathbf{n}_{c_2 k}$$

It is easy to show that for any set of coding data, $A_e^\pi \geq A_e^\kappa$ and therefore $\pi \leq \kappa$, with the limiting case (equality) obtaining when the observed distributions of the two coders are identical. The relationship between κ and S is not fixed.

2.5 More Than Two Coders

In corpus annotation practice, measuring reliability with only two coders is seldom considered enough, except for small-scale studies. Sometimes researchers run reliability studies with more than two coders, measure agreement separately for each pair of coders, and report the average. However, a better practice is to use generalized versions of the coefficients. A generalization of Scott's π is proposed in Fleiss (1971), and a generalization of Cohen's κ is given in Davies and Fleiss (1982). We will call these coefficients multi-π and multi-κ, respectively, dropping the multi-prefixes when no confusion is expected to arise.[5]

2.5.1 Fleiss's Multi-π. With more than two coders, the observed agreement A_o can no longer be defined as the percentage of items on which there is agreement, because inevitably there will be items on which some coders agree and others disagree. The solution proposed in the literature is to measure **pairwise agreement** (Fleiss 1971): Define the amount of agreement on a particular item as the proportion of agreeing judgment pairs out of the total number of judgment pairs for that item.

Multiple coders also pose a problem for the visualization of the data. When the number of coders $\mathbf{c}$ is greater than two, judgments cannot be shown in a contingency table like Table 1, because each coder has to be represented in a separate dimension.

5 Due to historical accident, the terminology in the literature is confusing. Fleiss (1971) proposed a coefficient of agreement for multiple coders and called it κ, even though it calculates expected agreement based on the cumulative distribution of judgments by all coders and is thus better thought of as a generalization of Scott's π. This unfortunate choice of name was the cause of much confusion in subsequent literature: Often, studies which claim to give a generalization of κ to more than two coders actually report Fleiss's coefficient (e.g., Bartko and Carpenter 1976; Siegel and Castellan 1988; Di Eugenio and Glass 2004). Since Carletta (1996) introduced reliability to the CL community based on the definitions of Siegel and Castellan (1988), the term "kappa" has been usually associated in this community with Siegel and Castellan's K, which is in effect Fleiss's coefficient, that is, a generalization of Scott's π.

Fleiss (1971) therefore uses a different type of table which lists each item with the number of judgments it received for each category; Siegel and Castellan (1988) use a similar table, which Di Eugenio and Glass (2004) call an **agreement table**. Table 3 is an example of an agreement table, in which the same 100 utterances from Table 1 are labeled by three coders instead of two. Di Eugenio and Glass (page 97) note that compared to contingency tables like Table 1, agreement tables like Table 3 lose information because they do not say which coder gave each judgment. This information is not used in the calculation of π, but is necessary for determining the individual coders' distributions in the calculation of κ. (Agreement tables also add information compared to contingency tables, namely, the identity of the items that make up each contingency class, but this information is not used in the calculation of either κ or π.)

Let $\mathbf{n}_{ik}$ stand for the number of times an item i is classified in category k (i.e., the number of coders that make such a judgment): For example, given the distribution in Table 3, $\mathbf{n}_{\text{Utt}_1\text{Stat}} = 2$ and $\mathbf{n}_{\text{Utt}_1\text{IReq}} = 1$. Each category k contributes $\binom{\mathbf{n}_{ik}}{2}$ pairs of agreeing judgments for item i; the amount of agreement agr_i for item i is the sum of $\binom{\mathbf{n}_{ik}}{2}$ over all categories $k \in K$, divided by $\binom{\mathbf{c}}{2}$, the total number of judgment pairs per item.

$$\text{agr}_i = \frac{1}{\binom{\mathbf{c}}{2}} \sum_{k \in K} \binom{\mathbf{n}_{ik}}{2} = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

For example, given the results in Table 3, we find the agreement value for Utterance 1 as follows.

$$\text{agr}_1 = \frac{1}{\binom{3}{2}} \left(\binom{\mathbf{n}_{\text{Utt}_1\text{Stat}}}{2} + \binom{\mathbf{n}_{\text{Utt}_1\text{IReq}}}{2} \right) = \frac{1}{3}(1+0) \approx 0.33$$

The overall observed agreement is the mean of agr_i for all items $i \in I$.

$$\text{A}_\text{o} = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i = \frac{1}{\mathbf{i}\mathbf{c}(\mathbf{c}-1)} \sum_{i \in I} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

(Notice that this definition of observed agreement is equivalent to the mean of the two-coder observed agreement values from Section 2.4 for all coder pairs.)

If observed agreement is measured on the basis of pairwise agreement (the proportion of agreeing judgment pairs), it makes sense to measure expected agreement in terms of pairwise comparisons as well, that is, as the probability that any pair of judgments for an item would be in agreement—or, said otherwise, the probability that two

Table 3
Agreement table with three coders.

	STAT	IREQ
Utt_1	2	1
Utt_2	0	3
⋮		
Utt_{100}	1	2
TOTAL	90 (0.3)	210 (0.7)

arbitrary coders would make the same judgment for a particular item by chance. This is the approach taken by Fleiss (1971). Like Scott, Fleiss interprets "chance agreement" as the agreement expected on the basis of a single distribution which reflects the combined judgments of all coders, meaning that expected agreement is calculated using $\hat{P}(k)$, the overall proportion of items assigned to category k, which is the total number of such assignments by all coders $\mathbf{n}_k$ divided by the overall number of assignments. The latter, in turn, is the number of items $\mathbf{i}$ multiplied by the number of coders $\mathbf{c}$.

$$\hat{P}(k) = \frac{1}{\mathbf{ic}} \mathbf{n}_k$$

As in the two-coder case, the probability that two arbitrary coders assign an item to a particular category $k \in K$ is assumed to be the joint probability of each coder making this assignment independently, that is $(\hat{P}(k))^2$. The expected agreement is the sum of this joint probability over all the categories $k \in K$.

$$A_e^{\pi} = \sum_{k \in K} \left(\hat{P}(k)\right)^2 = \sum_{k \in K} \left(\frac{1}{\mathbf{ic}} \mathbf{n}_k\right)^2 = \frac{1}{(\mathbf{ic})^2} \sum_{k \in K} \mathbf{n}_k^2$$

Multi-π is the coefficient that Siegel and Castellan (1988) call K.

2.5.2 Multi-κ. It is fairly straightforward to adapt Fleiss's proposal to generalize Cohen's κ proper to more than two coders, calculating expected agreement based on individual coder marginals. A detailed proposal can be found in Davies and Fleiss (1982), or in the extended version of this article.

2.6 Krippendorff's α and Other Weighted Agreement Coefficients

A serious limitation of both π and κ is that all disagreements are treated equally. But especially for semantic and pragmatic features, disagreements are not all alike. Even for the relatively simple case of dialogue act tagging, a disagreement between an `accept` and a `reject` interpretation of an utterance is clearly more serious than a disagreement between an `info-request` and a `check`. For tasks such as anaphora resolution, where reliability is determined by measuring agreement on sets (coreference chains), allowing for degrees of disagreement becomes essential (see Section 4.4). Under such circumstances, π and κ are not very useful.

In this section we discuss two coefficients that make it possible to differentiate between types of disagreements: α (Krippendorff 1980, 2004a), which is a coefficient defined in a general way that is appropriate for use with multiple coders, different magnitudes of disagreement, and missing values, and is based on assumptions similar to those of π; and weighted kappa κ_w (Cohen 1968), a generalization of κ.

2.6.1 Krippendorff's α. The coefficient α (Krippendorff 1980, 2004a) is an extremely versatile agreement coefficient based on assumptions similar to π, namely, that expected agreement is calculated by looking at the overall distribution of judgments without regard to which coders produced these judgments. It applies to multiple coders, and it allows for different magnitudes of disagreement. When all disagreements are considered equal it is nearly identical to multi-π, correcting for small sample sizes by using an unbiased estimator for expected agreement. In this section we will present

Krippendorff's α and relate it to the other coefficients discussed in this article, but we will start with α's origins as a measure of variance, following a long tradition of using variance to measure reliability (see citations in Rajaratnam 1960; Krippendorff 1970).

A sample's variance s^2 is defined as the sum of square differences from the mean $SS = \sum(x - \bar{x})^2$ divided by the degrees of freedom df. Variance is a useful way of looking at agreement if coders assign numerical values to the items, as in magnitude estimation tasks. Each item in a reliability study can be considered a separate level in a single-factor analysis of variance: The smaller the variance around each level, the higher the reliability. When agreement is perfect, the variance within the levels (s^2_{within}) is zero; when agreement is at chance, the variance within the levels is equal to the variance between the levels, in which case it is also equal to the overall variance of the data: $s^2_{within} = s^2_{between} = s^2_{total}$. The ratios $s^2_{within}/s^2_{between}$ (that is, $1/F$) and s^2_{within}/s^2_{total} are therefore 0 when agreement is perfect and 1 when agreement is at chance. Additionally, the latter ratio is bounded at 2: $SS_{within} \leq SS_{total}$ by definition, and $df_{total} < 2\,df_{within}$ because each item has at least two judgments. Subtracting the ratio s^2_{within}/s^2_{total} from 1 yields a coefficient which ranges between -1 and 1, where 1 signifies perfect agreement and 0 signifies chance agreement.

$$\alpha = 1 - \frac{s^2_{within}}{s^2_{total}} = 1 - \frac{SS_{within}/df_{within}}{SS_{total}/df_{total}}$$

We can unpack the formula for α to bring it to a form which is similar to the other coefficients we have looked at, and which will allow generalizing α beyond simple numerical values. The first step is to get rid of the notion of arithmetic mean which lies at the heart of the measure of variance. We observe that for any set of numbers $x_1, \ldots, x_N$ with a mean $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$, the sum of square differences from the mean SS can be expressed as the sum of square of differences between all the (ordered) pairs of numbers, scaled by a factor of $1/2N$.

$$SS = \sum_{n=1}^{N}(x_n - \bar{x})^2 = \frac{1}{2N}\sum_{n=1}^{N}\sum_{m=1}^{N}(x_n - x_m)^2$$

For calculating α we considered each item to be a separate level in an analysis of variance; the number of levels is thus the number of items $\mathbf{i}$, and because each coder marks each item, the number of observations for each item is the number of coders $\mathbf{c}$. Within-level variance is the sum of the square differences from the mean of each item, $SS_{within} = \sum_i \sum_c (x_{ic} - \bar{x}_i)^2$, divided by the degrees of freedom $df_{within} = \mathbf{i}(\mathbf{c} - 1)$. We can express this as the sum of the squares of the differences between all of the judgment pairs for each item, summed over all items and scaled by the appropriate factor. We use the notation x_{ic} for the value given by coder c to item i, and $\bar{x}_i$ for the mean of all the values given to item i.

$$s^2_{within} = \frac{SS_{within}}{df_{within}} = \frac{1}{\mathbf{i}(\mathbf{c}-1)}\sum_{i\in I}\sum_{c\in C}(x_{ic} - \bar{x}_i)^2 = \frac{1}{2\mathbf{ic}(\mathbf{c}-1)}\sum_{i\in I}\sum_{m=1}^{\mathbf{c}}\sum_{n=1}^{\mathbf{c}}(x_{ic_m} - x_{ic_n})^2$$

The total variance is the sum of the square differences of all judgments from the grand mean, $SS_{total} = \sum_i \sum_c (x_{ic} - \bar{x})^2$, divided by the degrees of freedom $df_{total} = \mathbf{ic} - 1$. This

can be expressed as the sum of the squares of the differences between all of the judgments pairs without regard to items, again scaled by the appropriate factor. The notation $\bar{x}$ is the overall mean of all the judgments in the data.

$$s^2_{total} = \frac{SS_{total}}{df_{total}} = \frac{1}{\mathbf{ic}-1}\sum_{i\in I}\sum_{c\in C}(x_{ic}-\bar{x})^2 = \frac{1}{2\mathbf{ic}(\mathbf{ic}-1)}\sum_{j=1}^{\mathbf{i}}\sum_{m=1}^{\mathbf{c}}\sum_{l=1}^{\mathbf{i}}\sum_{n=1}^{\mathbf{c}}(x_{i_jc_m}-x_{i_lc_n})^2$$

Now that we have removed references to means from our formulas, we can abstract over the measure of variance. We define a distance function $\mathbf{d}$ which takes two numbers and returns the square of their difference.

$$\mathbf{d}_{ab} = (a-b)^2$$

We also simplify the computation by counting all the identical value assignments together. Each unique value used by the coders will be considered a category $k \in K$. We use $\mathbf{n}_{ik}$ for the number of times item i is given the value k, that is, the number of coders that make such a judgment. For every (ordered) pair of distinct values $k_a, k_b \in K$ there are $\mathbf{n}_{ik_a}\mathbf{n}_{ik_b}$ pairs of judgments of item i, whereas for non-distinct values there are $\mathbf{n}_{ik_a}(\mathbf{n}_{ik_a}-1)$ pairs. We use this notation to rewrite the formula for the within-level variance. $\mathrm{D}^{\alpha}_{\mathrm{o}}$, the observed disagreement for α, is defined as twice the variance within the levels in order to get rid of the factor 2 in the denominator; we also simplify the formula by using the multiplier $\mathbf{n}_{ik_a}\mathbf{n}_{ik_a}$ for identical categories—this is allowed because $\mathbf{d}_{kk} = 0$ for all k.

$$\mathrm{D}^{\alpha}_{\mathrm{o}} = 2\,s^2_{within} = \frac{1}{\mathbf{ic}(\mathbf{c}-1)}\sum_{i\in I}\sum_{j=1}^{\mathbf{k}}\sum_{l=1}^{\mathbf{k}}\mathbf{n}_{ik_j}\mathbf{n}_{ik_l}\mathbf{d}_{k_jk_l}$$

We perform the same simplification for the total variance, where $\mathbf{n}_k$ stands for the total number of times the value k is assigned to any item by any coder. The expected disagreement for α, $\mathrm{D}^{\alpha}_{\mathrm{e}}$, is twice the total variance.

$$\mathrm{D}^{\alpha}_{\mathrm{e}} = 2\,s^2_{total} = \frac{1}{\mathbf{ic}(\mathbf{ic}-1)}\sum_{j=1}^{\mathbf{k}}\sum_{l=1}^{\mathbf{k}}\mathbf{n}_{k_j}\mathbf{n}_{k_l}\mathbf{d}_{k_jk_l}$$

Because both expected and observed disagreement are twice the respective variances, the coefficient α retains the same form when expressed with the disagreement values.

$$\alpha = 1 - \frac{\mathrm{D_o}}{\mathrm{D_e}}$$

Now that α has been expressed without explicit reference to means, differences, and squares, it can be generalized to a variety of coding schemes in which the labels cannot be interpreted as numerical values: All one has to do is to replace the square difference function $\mathbf{d}$ with a different distance function. Krippendorff (1980, 2004a) offers distance metrics suitable for nominal, interval, ordinal, and ratio scales. Of particular interest is

the function for nominal categories, that is, a function which considers all distinct labels equally distant from one another.

$$\mathbf{d}_{ab} = \begin{cases} 0 \text{ if } a = b \\ 1 \text{ if } a \neq b \end{cases}$$

It turns out that with this distance function, the observed disagreement D_o^α is exactly the complement of the observed agreement of Fleiss's multi-π, $1 - A_o^\pi$, and the expected disagreement D_e^α differs from $1 - A_e^\pi$ by a factor of $(\mathbf{ic} - 1)/\mathbf{ic}$; the difference is due to the fact that π uses a biased estimator of the expected agreement in the population whereas α uses an unbiased estimator. The following equation shows that given the correspondence between observed and expected agreement and disagreement, the coefficients themselves are nearly equivalent.

$$\alpha = 1 - \frac{D_o^\alpha}{D_e^\alpha} \approx 1 - \frac{1 - A_o^\pi}{1 - A_e^\pi} = \frac{1 - A_e^\pi - (1 - A_o^\pi)}{1 - A_e^\pi} = \frac{A_o^\pi - A_e^\pi}{1 - A_e^\pi} = \pi$$

For nominal data, the coefficients π and α approach each other as either the number of items or the number of coders approaches infinity.

Krippendorff's α will work with any distance metric, provided that identical categories always have a distance of zero ($\mathbf{d}_{kk} = 0$ for all k). Another useful constraint is symmetry ($\mathbf{d}_{ab} = \mathbf{d}_{ba}$ for all a, b). This flexibility affords new possibilities for analysis, which we will illustrate in Section 4. We should also note, however, that the flexibility also creates new pitfalls, especially in cases where it is not clear what the natural distance metric is. For example, there are different ways to measure dissimilarity between sets, and any of these measures can be justifiably used when the category labels are sets of items (as in the annotation of anaphoric relations). The different distance metrics yield different values of α for the same annotation data, making it difficult to interpret the resulting values. We will return to this problem in Section 4.4.

2.6.2 Cohen's κ_w. A weighted variant of Cohen's κ is presented in Cohen (1968). The implementation of weights is similar to that of Krippendorff's α—each pair of categories $k_a, k_b \in K$ is associated with a weight $\mathbf{d}_{k_a k_b}$, where a larger weight indicates more disagreement (Cohen uses the notation $\mathbf{v}$; he does not place any general constraints on the weights—not even a requirement that a pair of identical categories have a weight of zero, or that the weights be symmetric across the diagonal). The coefficient is defined for two coders: The disagreement for a particular item i is the weight of the pair of categories assigned to it by the two coders, and the overall observed disagreement is the (normalized) mean disagreement of all the items. Let $k(c_n, i)$ denote the category assigned by coder c_n to item i; then the disagreement for item i is $\text{disagr}_i = \mathbf{d}_{k(c_1,i)k(c_2,i)}$. The observed disagreement D_o is the mean of disagr_i for all items i, normalized to the interval $[0, 1]$ through division by the maximal weight $\mathbf{d}_{\max}$.

$$D_o^{\kappa_w} = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}} \sum_{i \in I} \text{disagr}_i = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}} \sum_{i \in I} \mathbf{d}_{k(c_1,i)k(c_2,i)}$$

If we take all disagreements to be of equal weight, that is $\mathbf{d}_{k_a k_a} = 0$ for all categories k_a and $\mathbf{d}_{k_a k_b} = 1$ for all $k_a \neq k_b$, then the observed disagreement is exactly the complement of the observed agreement as calculated in Section 2.4: $D_o^{\kappa_w} = 1 - A_o^\kappa$.

Like κ, the coefficient κ_w interprets expected disagreement as the amount expected by chance from a distinct probability distribution for each coder. These individual distributions are estimated by $\hat{\mathrm{P}}(k|c)$, the proportion of items assigned by coder c to category k, that is the number of such assignments $\mathbf{n}_{ck}$ divided by the number of items $\mathbf{i}$.

$$\hat{\mathrm{P}}(k|c) = \frac{1}{\mathbf{i}}\,\mathbf{n}_{ck}$$

The probability that coder c_1 assigns an item to category k_a and coder c_2 assigns it to category k_b is the joint probability of each coder making this assignment independently, namely, $\hat{\mathrm{P}}(k_a|c_1)\hat{\mathrm{P}}(k_b|c_2)$. The expected disagreement is the mean of the weights for all (ordered) category pairs, weighted by the probabilities of the category pairs and normalized to the interval $[0, 1]$ through division by the maximal weight.

$$\mathrm{D}_{\mathrm{e}}^{\kappa_w} = \frac{1}{\mathbf{d}_{\max}} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \hat{\mathrm{P}}(k_j|c_1)\hat{\mathrm{P}}(k_l|c_2)\mathbf{d}_{k_j k_l} = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}^2} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \mathbf{n}_{c_1 k_j} \mathbf{n}_{c_2 k_l} \mathbf{d}_{k_j k_l}$$

If we take all disagreements to be of equal weight then the expected disagreement is exactly the complement of the expected agreement for κ as calculated in Section 2.4: $\mathrm{D}_{\mathrm{e}}^{\kappa_w} = 1 - \mathrm{A}_{\mathrm{e}}^{\kappa}$.

Finally, the coefficient κ_w itself is the ratio of observed disagreement to expected disagreement, subtracted from 1 in order to yield a final value in terms of agreement.

$$\kappa_w = 1 - \frac{\mathrm{D_o}}{\mathrm{D_e}}$$

2.7 An Integrated Example

We end this section with an example illustrating how all of the agreement coefficients just discussed are computed. To facilitate comparisons, all computations will be based on the annotation statistics in Table 4. This confusion matrix reports the results of an experiment where two coders classify a set of utterances into three categories.

2.7.1 The Unweighted Coefficients. Observed agreement for all of the unweighted coefficients (S, κ, and π) is calculated by counting the items on which the coders agree (the

Table 4
An integrated coding example.

		Coder A			
		Stat	IReq	Chck	Total
Coder B	Stat	46	6	0	52
	IReq	0	32	0	32
	Chck	0	6	10	16
	Total	46	44	10	100

figures on the diagonal of the confusion matrix in Table 4) and dividing by the total number of items.

$$A_o = \frac{46 + 32 + 10}{100} = 0.88$$

The expected agreement values and the resulting values for the coefficients are shown in Table 5. The values of π and κ are very similar, which is to be expected when agreement is high, because this implies similar marginals. Notice that $A_e^\kappa < A_e^\pi$, hence $\kappa > \pi$; this reflects a general property of κ and π, already mentioned in Section 2.4, which will be elaborated in Section 3.1.

2.7.2 Weighted Coefficients. Suppose we notice that whereas `Statement` and `Info-Request` are clearly distinct classifications, `Check` is somewhere between the two. We therefore opt to weigh the distances between the categories as follows (recall that 1 denotes maximal disagreement, and identical categories are in full agreement and thus have a distance of 0).

	Statement	Info-Request	Check
Statement	0	1	0.5
Info-Request	1	0	0.5
Check	0.5	0.5	0

The observed disagreement is calculated by summing up *all* the cells in the contingency table, multiplying each cell by its respective weight, and dividing the total by the number of items (in the following calculation we ignore cells with zero items).

$$D_o = \frac{46 \times 0 + 6 \times 1 + 32 \times 0 + 6 \times 0.5 + 10 \times 0}{100} = \frac{6 + 3}{100} = 0.09$$

The only sources of disagreement in the coding example of Table 4 are the six utterances marked as `Info-Requests` by coder A and `Statements` by coder B, which receive the maximal weight of 1, and the six utterances marked as `Info-Requests` by coder A and `Checks` by coder B, which are given a weight of 0.5.

The calculation of expected disagreement for the weighted coefficients is shown in Table 6, and is the sum of the expected disagreement for each category pair multiplied

Table 5
Unweighted coefficients for the data from Table 4.

	Expected agreement	Chance-corrected agreement
S	$3 \times (\frac{1}{3})^2 = \frac{1}{3}$	$(0.88 - \frac{1}{3})/(1 - \frac{1}{3}) = 0.82$
π	$\frac{0.46+0.52}{2} + \frac{0.44+0.32}{2} + \frac{0.10+0.16}{2} = 0.4014$	$(0.88 - 0.4014)/(1 - 0.4014) \approx 0.7995$
κ	$.46 \times .52 + .44 \times .32 + .1 \times .16 = 0.396$	$(0.88 - 0.396)/(1 - 0.396) \approx 0.8013$

Table 6
Expected disagreement of the weighted coefficients for the data from Table 4.

D_e^{α}	$\frac{(46+52)\times(46+52)}{2\times100\times(2\times100-1)} \times 0 + \frac{(44+32)\times(46+52)}{2\times100\times(2\times100-1)} \times 1 + \frac{(10+16)\times(46+52)}{2\times100\times(2\times100-1)} \times \frac{1}{2}$ $+ \frac{(46+52)\times(44+32)}{2\times100\times(2\times100-1)} \times 1 + \frac{(44+32)\times(44+32)}{2\times100\times(2\times100-1)} \times 0 + \frac{(10+16)\times(44+32)}{2\times100\times(2\times100-1)} \times \frac{1}{2}$ $+ \frac{(46+52)\times(10+16)}{2\times100\times(2\times100-1)} \times \frac{1}{2} + \frac{(44+32)\times(10+16)}{2\times100\times(2\times100-1)} \times \frac{1}{2} + \frac{(10+16)\times(10+16)}{2\times100\times(2\times100-1)} \times 0$	0.4879
$D_e^{\kappa_w}$	$\frac{46\times52}{100\times100} \times 0 + \frac{44\times52}{100\times100} \times 1 + \frac{10\times52}{100\times100} \times \frac{1}{2}$ $+ \frac{46\times32}{100\times100} \times 1 + \frac{44\times32}{100\times100} \times 0 + \frac{10\times32}{100\times100} \times \frac{1}{2}$ $+ \frac{46\times16}{100\times100} \times \frac{1}{2} + \frac{44\times16}{100\times100} \times \frac{1}{2} + \frac{10\times16}{100\times100} \times 0$	0.49

by its weight. The value of the weighted coefficients is given by the formula $1 - \frac{D_o}{D_e}$, so $\alpha \approx 1 - \frac{0.09}{0.4879} \approx 0.8156$, and $\kappa_w = 1 - \frac{0.09}{0.49} \approx 0.8163$.

3. Bias and Prevalence

Two issues recently raised by Di Eugenio and Glass (2004) concern the behavior of agreement coefficients when the annotation data are severely skewed. One issue, which Di Eugenio and Glass call the **bias problem**, is that π and κ yield quite different numerical values when the annotators' marginal distributions are widely divergent; the other issue, the **prevalence problem**, is the exceeding difficulty in getting high agreement values when most of the items fall under one category. Looking at these two problems in detail is useful for understanding the differences between the coefficients.

3.1 Annotator Bias

The difference between π and α on the one hand and κ on the other hand lies in the interpretation of the notion of chance agreement, whether it is the amount expected from the the actual distribution of items among categories (π) or from individual coder priors (κ). As mentioned in Section 2.4, this difference has been the subject of much debate (Fleiss 1975; Krippendorff 1978, 2004b; Byrt, Bishop, and Carlin 1993; Zwick 1988; Hsu and Field 2003; Di Eugenio and Glass 2004; Craggs and McGee Wood 2005).

A claim often repeated in the literature is that single-distribution coefficients like π and α assume that different coders produce similar distributions of items among categories, with the implication that these coefficients are inapplicable when the annotators show substantially different distributions. Recommendations vary: Zwick (1988) suggests testing the individual coders' distributions using the modified χ^2 test of Stuart (1955), and discarding the annotation as unreliable if significant systematic discrepancies are observed. In contrast, Hsu and Field (2003, page 214) recommend reporting the value of κ even when the coders produce different distributions, because it is "the only [index] ... that could legitimately be applied in the presence of marginal heterogeneity"; likewise, Di Eugenio and Glass (2004, page 96) recommend using κ in "the vast majority ... of discourse- and dialogue-tagging efforts" where the individual coders' distributions tend to vary. All of these proposals are based on a misconception: that

single-distribution coefficients require similar distributions by the individual annotators in order to work properly. This is not the case. The difference between the coefficients is only in the interpretation of "chance agreement": π-style coefficients calculate the chance of agreement among *arbitrary* coders, whereas κ-style coefficients calculate the chance of agreement among the coders who produced the reliability data. Therefore, the choice of coefficient should not depend on the magnitude of the divergence between the coders, but rather on the desired interpretation of chance agreement.

Another common claim is that individual-distribution coefficients like κ "reward" annotators for disagreeing on the marginal distributions. For example, Di Eugenio and Glass (2004, page 99) say that κ suffers from what they call the bias problem, described as "the paradox that κ_{Co} [our κ] increases as the coders become less similar." Similar reservations about the use of κ have been noted by Brennan and Prediger (1981) and Zwick (1988). However, the bias problem is less paradoxical than it sounds. Although it is true that for a fixed observed agreement, a higher difference in coder marginals implies a lower expected agreement and therefore a higher κ value, the conclusion that κ penalizes coders for having similar distributions is unwarranted. This is because A_o and A_e are not independent: Both are drawn from the same set of observations. What κ does is discount some of the disagreement resulting from different coder marginals by incorporating it into A_e. Whether this is desirable depends on the application for which the coefficient is used.

The most common application of agreement measures in CL is to infer the reliability of a large-scale annotation, where typically each piece of data will be marked by just one coder, by measuring agreement on a small subset of the data which is annotated by multiple coders. In order to make this generalization, the measure must reflect the reliability of the annotation *procedure*, which is independent of the actual annotators used. Reliability, or reproducibility of the coding, is reduced by all disagreements—both random and systematic. The most appropriate measures of reliability for this purpose are therefore single-distribution coefficients like π and α, which generalize over the individual coders and exclude marginal disagreements from the expected agreement. This argument has been presented recently in much detail by Krippendorff (2004b) and reiterated by Craggs and McGee Wood (2005).

At the same time, individual-distribution coefficients like κ provide important information regarding the trustworthiness (validity) of the data on which the annotators agree. As an intuitive example, think of a person who consults two analysts when deciding whether to buy or sell certain stocks. If one analyst is an optimist and tends to recommend buying whereas the other is a pessimist and tends to recommend selling, they are likely to agree with each other less than two more neutral analysts, so overall their recommendations are likely to be less reliable—less reproducible—than those that come from a population of like-minded analysts. This reproducibility is measured by π. But whenever the optimistic and pessimistic analysts agree on a recommendation for a particular stock, whether it is "buy" or "sell," the confidence that this is indeed the right decision is higher than the same advice from two like-minded analysts. This is why κ "rewards" biased annotators: it is not a matter of reproducibility (reliability) but rather of trustworthiness (validity).

Having said this, we should point out that, first, in practice the difference between π and κ doesn't often amount to much (see discussion in Section 4). Moreover, the difference becomes smaller as agreement increases, because all the points of agreement contribute toward making the coder marginals similar (it took a lot of experimentation to create data for Table 4 so that the values of π and κ would straddle the conventional cutoff point of 0.80, and even so the difference is very small). Finally, one would expect

the difference between π and κ to diminish as the number of coders grows; this is shown subsequently.[6]

We define B, the overall **annotator bias** in a particular set of coding data, as the difference between the expected agreement according to (multi-)π and the expected agreement according to (multi-)κ. Annotator bias is a measure of variance: If we take c to be a random variable with equal probabilities for all coders, then the annotator bias B is the sum of the variances of $\mathrm{P}(k|c)$ for all categories $k \in K$, divided by the number of coders $\mathbf{c}$ less one (see Artstein and Poesio [2005] for a proof).

$$\mathrm{B} = \mathrm{A}_\mathrm{e}^\pi - \mathrm{A}_\mathrm{e}^\kappa = \frac{1}{\mathbf{c}-1} \sum_{k \in K} \sigma^2_{\hat{\mathrm{P}}(k|c)}$$

Annotator bias can be used to express the difference between κ and π.

$$\kappa - \pi = \frac{\mathrm{A_o} - (\mathrm{A}_\mathrm{e}^\pi - \mathrm{B})}{1 - (\mathrm{A}_\mathrm{e}^\pi - \mathrm{B})} - \frac{\mathrm{A_o} - \mathrm{A}_\mathrm{e}^\pi}{1 - \mathrm{A}_\mathrm{e}^\pi} = \mathrm{B} \cdot \frac{(1 - \mathrm{A_o})}{(1 - \mathrm{A}_\mathrm{e}^\kappa)(1 - \mathrm{A}_\mathrm{e}^\pi)}$$

This allows us to make the following observations about the relationship between π and κ.

Observation 1. *The difference between κ and π grows as the annotator bias grows: For a constant $\mathrm{A_o}$ and $\mathrm{A}_\mathrm{e}^\pi$, a greater* B *implies a greater value for $\kappa - \pi$.*

Observation 2. *The greater the number of coders, the lower the annotator bias* B*, and hence the lower the difference between κ and π, because the variance of $\hat{\mathrm{P}}(k|c)$ does not increase in proportion to the number of coders.*

In other words, provided enough coders are used, it should not matter whether a single-distribution or individual-distribution coefficient is used. This is not to imply that multiple coders increase reliability: The variance of the individual coders' distributions can be just as large with many coders as with few coders, but its effect on the value of κ decreases as the number of coders grows, and becomes more similar to random noise.

The same holds for weighted measures too; see the extended version of this article for definitions and proof. In an annotation study with 18 subjects, we compared α with a variant which uses individual coder distributions to calculate expected agreement, and found that the values never differed beyond the third decimal point (Poesio and Artstein 2005).

We conclude with a summary of our views concerning the difference between π-style and κ-style coefficients. First of all, keep in mind that empirically the difference is small, and gets smaller as the number of annotators increases. Then instead of reporting two coefficients, as suggested by Di Eugenio and Glass (2004), the appropriate coefficient should be chosen based on the task (*not* on the observed differences between coder marginals). When the coefficient is used to assess reliability, a single-distribution coefficient like π or α should be used; this is indeed already the practice in CL, because Siegel and Castellan's K is identical with (multi-)π. It is also good practice to test

6 Craggs and McGee Wood (2005) also suggest increasing the number of coders in order to overcome individual annotator bias, but do not provide a mathematical justification.

reliability with more than two coders, in order to reduce the likelihood of coders sharing a deviant reading of the annotation guidelines.

3.2 Prevalence

We touched upon the matter of skewed data in Section 2.3 when we motivated the need for chance correction: If a disproportionate amount of the data falls under one category, then the expected agreement is very high, so in order to demonstrate high reliability an even higher observed agreement is needed. This leads to the so-called paradox that chance-corrected agreement may be low even though A_o is high (Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990; Di Eugenio and Glass 2004). Moreover, when the data are highly skewed in favor of one category, the high agreement also corresponds to high accuracy: If, say, 95% of the data fall under one category label, then random coding would cause two coders to jointly assign this category label to 90.25% of the items, and on average 95% of these labels would be correct, for an overall accuracy of at least 85.7%. This leads to the surprising result that when data are highly skewed, coders may agree on a high proportion of items while producing annotations that are indeed correct to a high degree, yet the reliability coefficients remain low. (For an illustration, see the discussion of agreement results on coding discourse segments in Section 4.3.1.)

This surprising result is, however, justified. Reliability implies the ability to distinguish between categories, but when one category is very common, high accuracy and high agreement can also result from indiscriminate coding. The test for reliability in such cases is the ability to agree on the rare categories (regardless of whether these are the categories of interest). Indeed, chance-corrected coefficients are sensitive to agreement on rare categories. This is easiest to see with a simple example of two coders and two categories, one common and the other one rare; to further simplify the calculation we also assume that the coder marginals are identical, so that π and κ yield the same values. We can thus represent the judgments in a contingency table with just two parameters: ϵ is half the proportion of items on which there is disagreement, and δ is the proportion of agreement on the **Rare** category. Both of these proportions are assumed to be small, so the bulk of the items (a proportion of $1-(\delta+2\epsilon)$) are labeled with the **Common** category by both coders (Table 7). From this table we can calculate $A_o = 1-2\epsilon$ and $A_e = 1-2(\delta+\epsilon)+2(\delta+\epsilon)^2$, as well as π and κ.

$$\pi,\kappa = \frac{1-2\epsilon-(1-2(\delta+\epsilon)+2(\delta+\epsilon)^2)}{1-(1-2(\delta+\epsilon)+2(\delta+\epsilon)^2)} = \frac{\delta}{\delta+\epsilon} - \frac{\epsilon}{1-(\delta+\epsilon)}$$

When ϵ and δ are both small, the fraction after the minus sign is small as well, so π and κ are approximately $\delta/(\delta+\epsilon)$: the value we get if we take all the items marked by one

Table 7
A simple example of agreement on dialogue act tagging.

		CODER A		
		COMMON	RARE	TOTAL
CODER B	COMMON	$1-(\delta+2\epsilon)$	ϵ	$1-(\delta+\epsilon)$
	RARE	ϵ	δ	$\delta+\epsilon$
	TOTAL	$1-(\delta+\epsilon)$	$\delta+\epsilon$	1

particular coder as **Rare**, and calculate what proportion of those items were labeled **Rare** by the other coder. This is a measure of the coders' ability to agree on the rare category.

4. Using Agreement Measures for CL Annotation Tasks

In this section we review the use of intercoder agreement measures in CL since Carletta's original paper in light of the discussion in the previous sections. We begin with a summary of Krippendorff's recommendations about measuring reliability (Krippendorff 2004a, Chapter 11), then discuss how coefficients of agreement have been used in CL to measure the reliability of annotation schemes, focusing in particular on the types of annotation where there has been some debate concerning the most appropriate measures of agreement.

4.1 Methodology and Interpretation of the Results: General Issues

Krippendorff (2004a, Chapter 11) notes with regret the fact that reliability is discussed in only around 69% of studies in content analysis. In CL as well, not all annotation projects include a formal test of intercoder agreement. Some of the best known annotation efforts, such as the creation of the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993) and the British National Corpus (Leech, Garside, and Bryant 1994), do not report reliability results as they predate the Carletta paper; but even among the more recent efforts, many only report percentage agreement, as for the creation of the PropBank (Palmer, Dang, and Fellbaum 2007) or the ongoing OntoNotes annotation (Hovy et al. 2006). Even more importantly, very few studies apply a methodology as rigorous as that envisaged by Krippendorff and other content analysts. We therefore begin this discussion of CL practice with a summary of the main recommendations found in Chapter 11 of Krippendorff (2004a), even though, as we will see, we think that some of these recommendations may not be appropriate for CL.

4.1.1 Generating Data to Measure Reproducibility. Krippendorff's recommendations were developed for the field of content analysis, where coding is used to draw conclusions from the texts. A coded corpus is thus akin to the result of a scientific experiment, and it can only be considered valid if it is reproducible—that is, if the same coded results can be replicated in an independent coding exercise. Krippendorff therefore argues that any study using observed agreement as a measure of reproducibility must satisfy the following requirements:

- It must employ an exhaustively formulated, clear, and usable coding scheme together with step-by-step instructions on how to use it.
- It must use clearly specified criteria concerning the choice of coders (so that others may use such criteria to reproduce the data).
- It must ensure that the coders that generate the data used to measure reproducibility work independently of each other.

Some practices that are common in CL do not satisfy these requirements. The first requirement is violated by the practice of expanding the written coding instructions and including new rules as the data are generated. The second requirement is often

violated by using experts as coders, particularly long-term collaborators, as such coders may agree not because they are carefully following written instructions, but because they know the purpose of the research very well—which makes it virtually impossible for others to reproduce the results on the basis of the same coding scheme (the problems arising when using experts were already discussed at length in Carletta [1996]). Practices which violate the third requirement (independence) include asking coders to discuss their judgments with each other and reach their decisions by majority vote, or to consult with each other when problems not foreseen in the coding instructions arise. Any of these practices make the resulting data unusable for measuring reproducibility.

Krippendorff's own summary of his recommendations is that to obtain usable data for measuring reproducibility a researcher must use data generated by three or more coders, chosen according to some clearly specified criteria, and working independently according to a written coding scheme and coding instructions fixed in advance. Krippendorff also discusses the criteria to be used in the selection of the sample, from the minimum number of units (obtained using a formula from Bloch and Kraemer [1989], reported in Krippendorff [2004a, page 239]), to how to make the sample representative of the data population (each category should occur in the sample often enough to yield at least five chance agreements), to how to ensure the reliability of the instructions (the sample should contain examples of all the values for the categories). These recommendations are particularly relevant in light of the comments of Craggs and McGee Wood (2005, page 290), which discourage researchers from testing their coding instructions on data from more than one domain. Given that the reliability of the coding instructions depends to a great extent on how complications are dealt with, and that every domain displays different complications, the sample should contain sufficient examples from all domains which have to be annotated according to the instructions.

4.1.2 Establishing Significance. In hypothesis testing, it is common to test for the significance of a result against a null hypothesis of chance behavior; for an agreement coefficient this would mean rejecting the possibility that a positive value of agreement is nevertheless due to random coding. We can rely on the statement by Siegel and Castellan (1988, Section 9.8.2) that when sample sizes are large, the sampling distribution of K (Fleiss's multi-π) is approximately normal and centered around zero—this allows testing the obtained value of K against the null hypothesis of chance agreement by using the z statistic. It is also easy to test Krippendorff's α with the interval distance metric against the null hypothesis of chance agreement, because the hypothesis $\alpha = 0$ is identical to the hypothesis $F = 1$ in an analysis of variance.

However, a null hypothesis of chance agreement is not very interesting, and demonstrating that agreement is significantly better than chance is not enough to establish reliability. This has already been pointed out by Cohen (1960, page 44): "to know merely that κ is beyond chance is trivial since one usually expects much more than this in the way of reliability in psychological measurement." The same point has been repeated and stressed in many subsequent works (e.g., Posner et al. 1990; Di Eugenio 2000; Krippendorff 2004a): The reason for measuring reliability is not to test whether coders perform better than chance, but to ensure that the coders do not deviate too much from perfect agreement (Krippendorff 2004a, page 237).

The relevant notion of significance for agreement coefficients is therefore a confidence interval. Cohen (1960, pages 43–44) implies that when sample sizes are large, the sampling distribution of κ is approximately normal for any true population value of κ, and therefore confidence intervals for the observed value of κ can be determined

using the usual multiples of the standard error. Donner and Eliasziw (1987) propose a more general form of significance test for arbitrary levels of agreement. In contrast, Krippendorff (2004a, Section 11.4.2) states that the distribution of α is unknown, so confidence intervals must be obtained by bootstrapping; a software package for doing this is described in Hayes and Krippendorff (2007).

4.1.3 Interpreting the Value of Kappa-Like Coefficients. Even after testing significance and establishing confidence intervals for agreement coefficients, we are still faced with the problem of interpreting the meaning of the resulting values. Suppose, for example, we establish that for a particular task, $K = 0.78 \pm 0.05$. Is this good or bad? Unfortunately, deciding what counts as an adequate level of agreement for a specific purpose is still little more than a black art: As we will see, different levels of agreement may be appropriate for resource building and for more linguistic purposes.

The problem is not unlike that of interpreting the values of correlation coefficients, and in the area of medical diagnosis, the best known conventions concerning the value of kappa-like coefficients, those proposed by Landis and Koch (1977) and reported in Figure 1, are indeed similar to those used for correlation coefficients, where values above 0.4 are also generally considered adequate (Marion 2004). Many medical researchers feel that these conventions are appropriate, and in language studies, a similar interpretation of the values has been proposed by Rietveld and van Hout (1993). In CL, however, most researchers follow the more stringent conventions from content analysis proposed by Krippendorff (1980, page 147), as reported by Carletta (1996, page 252): "content analysis researchers generally think of $K > .8$ as good reliability, with $.67 < K < .8$ allowing tentative conclusions to be drawn" (Krippendorff was discussing values of α rather than K, but the coefficients are nearly equivalent for categorical labels). As a result, ever since Carletta's influential paper, CL researchers have attempted to achieve a value of K (more seldom, of α) above the 0.8 threshold, or, failing that, the 0.67 level allowing for "tentative conclusions." However, the description of the 0.67 boundary in Krippendorff (1980) was actually "highly tentative and cautious," and in later work Krippendorff clearly considers 0.8 the absolute minimum value of α to accept for any serious purpose: "Even a cutoff point of $\alpha = .800 \ldots$ is a pretty low standard" (Krippendorff 2004a, page 242). Recent content analysis practice seems to have settled for even more stringent requirements: A recent textbook, Neuendorf (2002, page 3), analyzing several proposals concerning "acceptable" reliability, concludes that "reliability coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations, and below that, there exists great disagreement."

This is clearly a fundamental issue. Ideally we would want to establish thresholds which are appropriate for the field of CL, but as we will see in the rest of this section, a decade of practical experience hasn't helped in settling the matter. In fact, weighted coefficients, while arguably more appropriate for many annotation tasks, make the issue of deciding when the value of a coefficient indicates sufficient agreement even

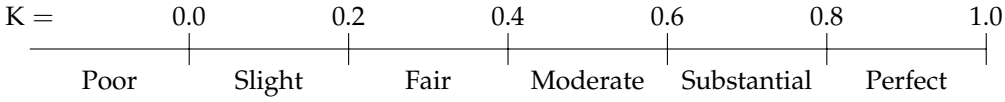

Figure 1
Kappa values and strength of agreement according to Landis and Koch (1977).

more complicated because of the problem of determining appropriate weights (see Section 4.4). We will return to the issue of interpreting the value of the coefficients at the end of this article.

4.1.4 Agreement and Machine Learning. In a recent article, Reidsma and Carletta (2008) point out that the goals of annotation in CL differ from those of content analysis, where agreement coefficients originate. A common use of an annotated corpus in CL is not to confirm or reject a hypothesis, but to generalize the patterns using machine-learning algorithms. Through a series of simulations, Reidsma and Carletta demonstrate that agreement coefficients are poor predictors of machine-learning success: Even highly reproducible annotations are difficult to generalize when the disagreements contain patterns that can be learned, whereas highly noisy and unreliable data can be generalized successfully when the disagreements do not contain learnable patterns. These results show that agreement coefficients should not be used as indicators of the suitability of annotated data for machine learning.

However, the purpose of reliability studies is not to find out whether annotations can be generalized, but whether they capture some kind of observable reality. Even if the pattern of disagreement allows generalization, we need evidence that this generalization would be meaningful. The decision whether a set of annotation guidelines are appropriate or meaningful is ultimately a qualitative one, but a baseline requirement is an acceptable level of agreement among the annotators, who serve as the instruments of measurement. Reliability studies test the soundness of an annotation scheme and guidelines, which is not to be equated with the machine-learnability of data produced by such guidelines.

4.2 Labeling Units with a Common and Predefined Set of Categories: The Case of Dialogue Act Tagging

The simplest and most common coding in CL involves labeling segments of text with a limited number of linguistic categories: Examples include part-of-speech tagging, dialogue act tagging, and named entity tagging. The practices used to test reliability for this type of annotation tend to be based on the assumption that the categories used in the annotation are mutually exclusive and equally distinct from one another; this assumption seems to have worked out well in practice, but questions about it have been raised even for the annotation of parts of speech (Babarczy, Carroll, and Sampson 2006), let alone for discourse coding tasks such as dialogue act coding. We concentrate here on this latter type of coding, but a discussion of issues raised for POS, named entity, and prosodic coding can be found in the extended version of the article.

Dialogue act tagging is a type of linguistic annotation with which by now the CL community has had extensive experience: Several dialogue-act-annotated spoken language corpora now exist, such as MapTask (Carletta et al. 1997), Switchboard (Stolcke et al. 2000), Verbmobil (Jekat et al. 1995), and Communicator (e.g., Doran et al. 2001), among others. Historically, dialogue act annotation was also one of the types of annotation that motivated the introduction in CL of chance-corrected coefficients of agreement (Carletta et al. 1997) and, as we will see, it has been the type of annotation that has generated the most discussion concerning annotation methodology and measuring agreement.

A number of coding schemes for dialogue acts have achieved values of K over 0.8 and have therefore been assumed to be reliable: For example, $K = 0.83$ for the

13-tag MapTask coding scheme (Carletta et al. 1997), K = 0.8 for the 42-tag Switchboard-DAMSL scheme (Stolcke et al. 2000), K = 0.90 for the smaller 20-tag subset of the CSTAR scheme used by Doran et al. (2001). All of these tests were based on the same two assumptions: that every unit (utterance) is assigned to exactly one category (dialogue act), and that these categories are distinct. Therefore, again, unweighted measures, and in particular K, tend to be used for measuring inter-coder agreement.

However, these assumptions have been challenged based on the observation that utterances tend to have more than one function at the dialogue act level (Traum and Hinkelman 1992; Allen and Core 1997; Bunt 2000); for a useful survey, see Popescu-Belis (2005). An assertion performed in answer to a question, for instance, typically performs at least two functions at different levels: asserting some information—the dialogue act that we called `Statement` in Section 2.3, operating at what Traum and Hinkelman called the "core speech act" level—and confirming that the question has been understood, a dialogue act operating at the "grounding" level and usually known as Acknowledgment (`Ack`). In older dialogue act tagsets, acknowledgments and statements were treated as alternative labels at the same "level", forcing coders to choose one or the other when an utterance performed a dual function, according to a well-specified set of instructions. By contrast, in the annotation schemes inspired from these newer theories such as DAMSL (Allen and Core 1997), coders are allowed to assign tags along distinct "dimensions" or "levels".

Two annotation experiments testing this solution to the "multi-tag" problem with the DAMSL scheme were reported in Core and Allen (1997) and Di Eugenio et al. (1998). In both studies, coders were allowed to mark each communicative function independently: That is, they were allowed to choose for each utterance one of the `Statement` tags (or possibly none), one of the `Influencing-Addressee-Future-Action` tags, and so forth—and agreement was evaluated separately for each dimension using (unweighted) K. Core and Allen found values of K ranging from 0.76 for `answer` to 0.42 for `agreement` to 0.15 for `Committing-Speaker-Future-Action`. Using different coding instructions and on a different corpus, Di Eugenio et al. observed higher agreement, ranging from K = 0.93 (for `other-forward-function`) to 0.54 (for the tag `agreement`).

These relatively low levels of agreement led many researchers to return to "flat" tagsets for dialogue acts, incorporating however in their schemes some of the insights motivating the work on schemes such as DAMSL. The best known example of this type of approach is the development of the SWITCHBOARD-DAMSL tagset by Jurafsky, Shriberg, and Biasca (1997), which incorporates many ideas from the "multi-dimensional" theories of dialogue acts, but does not allow marking an utterance as both an acknowledgment and a statement; a choice has to be made. This tagset results in overall agreement of K = 0.80. Interestingly, subsequent developments of SWITCHBOARD-DAMSL backtracked on some of these decisions. For instance, the ICSI-MRDA tagset developed for the annotation of the ICSI Meeting Recorder corpus reintroduces some of the DAMSL ideas, in that annotators are allowed to assign multiple SWITCHBOARD-DAMSL labels to utterances (Shriberg et al. 2004). Shriberg et al. achieved a comparable reliability to that obtained with SWITCHBOARD-DAMSL, but only when using a tagset of just five "class-maps".

Shriberg et al. (2004) also introduced a hierarchical organization of tags to improve reliability. The dimensions of the DAMSL scheme can be viewed as "superclasses" of dialogue acts which share some aspect of their meaning. For instance, the dimension of `Influencing-Addressee-Future-Action` (IAFA) includes the two dialogue acts `Open-option` (used to mark suggestions) and `Directive`, both of which bring into

consideration a future action to be performed by the addressee. At least in principle, an organization of this type opens up the possibility for coders to mark an utterance with the superclass (IAFA) in case they do not feel confident that the utterance satisfies the additional requirements for `Open-option` or `Directive`. This, in turn, would do away with the need to make a choice between these two options. This possibility wasn't pursued in the studies using the original DAMSL that we are aware of (Core and Allen 1997; Di Eugenio 2000; Stent 2001), but was tested by Shriberg et al. (2004) and subsequent work, in particular Geertzen and Bunt (2006), who were specifically interested in the idea of using hierarchical schemes to measure partial agreement, and in addition experimented with weighted coefficients of agreement for their hierarchical tagging scheme, specifically κ_w.

Geertzen and Bunt tested intercoder agreement with Bunt's DIT++ (Bunt 2005), a scheme with 11 dimensions that builds on ideas from DAMSL and from Dynamic Interpretation Theory (Bunt 2000). In DIT++, tags can be hierarchically related: For example, the class `information-seeking` is viewed as consisting of two classes, yes-no question (`ynq`) and *wh*-question (`whq`). The hierarchy is explicitly introduced in order to allow coders to leave some aspects of the coding undecided. For example, `check` is treated as a subclass of `ynq` in which, in addition, the speaker has a weak belief that the proposition that forms the belief is true. A coder who is not certain about the dialogue act performed using an utterance may simply choose to tag it as `ynq`.

The distance metric **d** proposed by Geertzen and Bunt is based on the criterion that two communicative functions are related ($\mathbf{d}(c_1, c_2) < 1$) if they stand in an ancestor–offspring relation within a hierarchy. Furthermore, they argue, the magnitude of $\mathbf{d}(c_1, c_2)$ should be proportional to the distance between the functions in the hierarchy. A level-dependent correction factor is also proposed so as to leave open the option to make disagreements at higher levels of the hierarchy matter more than disagreements at the deeper level (for example, the distance between `information-seeking` and `ynq` might be considered greater than the distance between `check` and `positive-check`).

The results of an agreement test with two annotators run by Geertzen and Bunt show that taking into account partial agreement leads to values of κ_w that are higher than the values of κ for the same categories, particularly for `feedback`, a class for which Core and Allen (1997) got low agreement. Of course, even assuming that the values of κ_w and κ were directly comparable—we remark on the difficulty of interpreting the values of weighted coefficients of agreement in Section 4.4—it remains to be seen whether these higher values are a better indication of the extent of agreement between coders than the values of unweighted κ.

This discussion of coding schemes for dialogue acts introduced issues to which we will return for other CL annotation tasks as well. There are a number of well-established schemes for large-scale dialogue act annotation based on the assumption of mutual exclusivity between dialogue act tags, whose reliability is also well known; if one of these schemes is appropriate for modeling the communicative intentions found in a task, we recommend to our readers to use it. They should also realize, however, that the mutual exclusivity assumption is somewhat dubious. If a multi-dimensional or hierarchical tagset is used, readers should also be aware that weighted coefficients do capture partial agreement, and need not automatically result in lower reliability or in an explosion in the number of labels. However, a hierarchical scheme may not reflect genuine annotation difficulties: For example, in the case of DIT++, one might argue that it is more difficult to confuse yes-no questions with *wh*-questions than with statements. We will also see in a moment that interpreting the results with weighted coefficients is difficult. We will return to both of these problems in what follows.

4.3 Marking Boundaries and Unitizing

Before labeling can take place, the units of annotation, or markables, need to be identified—a process Krippendorff (1995, 2004a) calls **unitizing**. The practice in CL for the forms of annotation discussed in the previous section is to assume that the units are linguistic constituents which can be easily identified, such as words, utterances, or noun phrases, and therefore there is no need to check the reliability of this process. We are aware of few exceptions to this assumption, such as Carletta et al. (1997) on unitization for move coding and our own work on the GNOME corpus (Poesio 2004b). In cases such as text segmentation, however, the identification of units is as important as their labeling, if not more important, and therefore checking agreement on unit identification is essential. In this section we discuss current CL practice with reliability testing of these types of annotation, before briefly summarizing Krippendorff's proposals concerning measuring reliability for unitizing.

4.3.1 Segmentation and Topic Marking. Discourse segments are portions of text that constitute a unit either because they are about the same "topic" (Hearst 1997; Reynar 1998) or because they have to do with achieving the same intention (Grosz and Sidner 1986) or performing the same "dialogue game" (Carletta et al. 1997).[7] The analysis of discourse structure—and especially the identification of discourse segments—is the type of annotation that, more than any other, led CL researchers to look for ways of measuring reliability and agreement, as it made them aware of the extent of disagreement on even quite simple judgments (Kowtko, Isard, and Doherty 1992; Passonneau and Litman 1993; Carletta et al. 1997; Hearst 1997). Subsequent research identified a number of issues with discourse structure annotation, above all the fact that segmentation, though problematic, is still much easier than marking more complex aspects of discourse structure, such as identifying the most important segments or the "rhetorical" relations between segments of different granularity. As a result, many efforts to annotate discourse structure concentrate only on segmentation.

The agreement results for segment coding tend to be on the lower end of the scale proposed by Krippendorff and recommended by Carletta. Hearst (1997), for instance, found $K = 0.647$ for the boundary/not boundary distinction; Reynar (1998), measuring agreement between his own annotation and the TREC segmentation of broadcast news, reports $K = 0.764$ for the same task; Ries (2002) reports even lower agreement of $K = 0.36$. Teufel, Carletta, and Moens (1999), who studied agreement on the identification of **argumentative zones**, found high reliability ($K = 0.81$) for their three main zones (own, other, background), although lower for the whole scheme ($K = 0.71$). For intention-based segmentation, Passonneau and Litman (1993) in the pre-K days reported an overall percentage agreement with majority opinion of 89%, but the agreement on boundaries was only 70%. For conversational games segmentation, Carletta et al. (1997) reported "promising but not entirely reassuring agreement on where games began (70%)," whereas the agreement on transaction boundaries was $K = 0.59$. Exceptions are two segmentation efforts carried out as part of annotations of rhetorical structure. Moser, Moore, and Glendening (1996) achieved an agreement

7 The notion of "topic" is notoriously difficult to define and many competing theoretical proposals exist (Reinhart 1981; Vallduví 1993). As it is often the case with annotation, fairly simple definitions tend to be used in discourse annotation work: For example, in TDT topic is defined for annotation purposes as "an event or activity, along with all directly related events and activities" (TDT-2 Annotation Guide, `http://projects.ldc.upenn.edu/TDT2/Guide/label-instr.html`).

of $K = 0.9$ for the highest level of segmentation of their RDA annotation (Poesio, Patel, and Di Eugenio 2006). Carlson, Marcu, and Okurowski (2003) reported very high agreement over the identification of the boundaries of discourse units, the building blocks of their annotation of rhetorical structure. (Agreement was measured several times; initially, they obtained $K = 0.87$, and in the final analysis $K = 0.97$.) This, however, was achieved by employing experienced annotators, and with considerable training.

One important reason why most agreement results on segmentation are on the lower end of the reliability scale is the fact, known to researchers in discourse analysis from as early as Levin and Moore (1978), that although analysts generally agree on the "bulk" of segments, they tend to disagree on their exact boundaries. This phenomenon was also observed in more recent studies: See for example the discussion in Passonneau and Litman (1997), the comparison of the annotations produced by seven coders of the same text in Figure 5 of Hearst (1997, page 55), or the discussion by Carlson, Marcu, and Okurowski (2003), who point out that the boundaries between elementary discourse units tend to be "very blurry." See also Pevzner and Hearst (2002) for similar comments made in the context of topic segmentation algorithms, and Klavans, Popper, and Passonneau (2003) for selecting definition phrases.

This "blurriness" of boundaries, combined with the prevalence effects discussed in Section 3.2, also explains the fact that topic annotation efforts which were only concerned with roughly dividing a text into segments (Passonneau and Litman 1993; Carletta et al. 1997; Hearst 1997; Reynar 1998; Ries 2002) generally report lower agreement than the studies whose goal is to identify smaller discourse units. When disagreement is mostly concentrated in one class ('boundary' in this case), if the total number of units to annotate remains the same, then expected agreement on this class is lower when a greater proportion of the units to annotate belongs to this class. When in addition this class is much less numerous than the other classes, overall agreement tends to depend mostly on agreement on this class.

For instance, suppose we are testing the reliability of two different segmentation schemes—into broad "discourse segments" and into finer "discourse units"—on a text of 50 utterances, and that we obtain the results in Table 8. Case 1 would be a situation in which Coder A and Coder B agree that the text consists of two segments, obviously agree on its initial and final boundaries, but disagree by one position on the intermediate boundary—say, one of them places it at utterance 25, the other at utterance 26. Nevertheless, because expected agreement is so high—the coders agree on the classification of 98% of the utterances—the value of K is fairly low. In case 2, the coders disagree on three times as many utterances, but K is higher than in the first case because expected agreement is substantially lower ($A_e = 0.53$).

The fact that coders mostly agree on the "bulk" of discourse segments, but tend to disagree on their boundaries, also makes it likely that an all-or-nothing coefficient like K calculated on individual boundaries would underestimate the degree of agreement, suggesting low agreement even among coders whose segmentations are mostly similar. A weighted coefficient of agreement like α might produce values more in keeping with intuition, but we are not aware of any attempts at measuring agreement on segmentation using weighted coefficients. We see two main options. We suspect that the methods proposed by Krippendorff (1995) for measuring agreement on unitizing (see Section 4.3.2, subsequently) may be appropriate for the purpose of measuring agreement on discourse segmentation. A second option would be to measure agreement not on individual boundaries but on windows spanning several units, as done in the methods proposed to evaluate the performance of topic detection algorithms such as

Table 8
Fewer boundaries, higher expected agreement.

Case 1: Broad segments $A_o = 0.96, A_e = 0.89, K = 0.65$				
		CODER A		
		BOUNDARY	NO BOUNDARY	TOTAL
CODER B	BOUNDARY	2	1	3
	NO BOUNDARY	1	46	47
	TOTAL	3	47	50

Case 2: Fine discourse units $A_o = 0.88, A_e = 0.53, K = 0.75$				
		CODER A		
		BOUNDARY	NO BOUNDARY	TOTAL
CODER B	BOUNDARY	16	3	19
	NO BOUNDARY	3	28	31
	TOTAL	19	31	50

P_k (Beeferman, Berger, and Lafferty 1999) or WINDOWDIFF (Pevzner and Hearst 2002) (which are, however, raw agreement scores not corrected for chance).

4.3.2 Unitizing (Or, Agreement on Markable Identification). It is often assumed in CL annotation practice that the units of analysis are "natural" linguistic objects, and therefore there is no need to check agreement on their identification. As a result, agreement is usually measured on the labeling of units rather than on the process of identifying them (**unitizing**, Krippendorff 1995). We have just seen, however, two coding tasks for which the reliability of unit identification is a crucial part of the overall reliability, and the problem of markable identification is more pervasive than is generally acknowledged. For example, when the units to be labeled are syntactic constituents, it is common practice to use a parser or chunker to identify the markables and then to allow the coders to correct the parser's output. In such cases one would want to know how reliable the coders' corrections are. We thus need a general method of testing relibility on markable identification.

The one proposal for measuring agreement on markable identification we are aware of is the α_U coefficient, a non-trivial variant of α proposed by Krippendorff (1995). A full presentation of the proposal would require too much space, so we will just present the core idea. Unitizing is conceived of as consisting of two separate steps: identifying boundaries between units, and selecting the units of interest. If a unit identified by one coder overlaps a unit identified by the other coder, the amount of disagreement is the square of the lengths of the non-overlapping segments (see Figure 2); if a unit identified by one coder does not overlap any unit of interest identified by the other coder, the amount of disagreement is the square of the length of the whole unit. This distance metric is used in calculating observed and expected disagreement, and α_U itself. We refer the reader to Krippendorff (1995) for details.

Krippendorff's α_U is not applicable to all CL tasks. For example, it assumes that units may not overlap in a single coder's output, yet in practice there are many

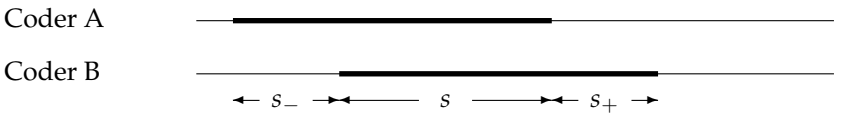

Figure 2
The difference between overlapping units is $\mathbf{d}(A, B) = s_-^2 + s_+^2$ (adapted from Krippendorff 1995, Figure 4, page 61).

annotation schemes which require coders to label nested syntactic constituents. For continuous segmentation tasks, α_U may be inappropriate because when a segment identified by one annotator overlaps with two segments identified by another annotator, the distance is smallest when the one segment is centered over the two rather than aligned with one of them. Nevertheless, we feel that when the non-overlap assumption holds, and the units do not cover the text exhaustively, testing the reliabilty of unit identification may prove beneficial. To our knowledge, this has never been tested in CL.

4.4 Anaphora

The annotation tasks discussed so far involve assigning a specific label to each category, which allows the various agreement measures to be applied in a straightforward way. Anaphoric annotation differs from the previous tasks because annotators do not assign labels, but rather create links between anaphors and their antecedents. It is therefore not clear what the "labels" should be for the purpose of calculating agreement. One possibility would be to consider the intended referent (real-world object) as the label, as in named entity tagging, but it wouldn't make sense to predefine a set of "labels" applicable to all texts, because different objects are mentioned in different texts. An alternative is to use the marked antecedents as "labels". However, we do not want to count as a disagreement every time two coders agree on the discourse entity realized by a particular noun phrase but just happen to mark different words as antecedents. Consider the reference of the underlined pronoun *it* in the following dialogue excerpt (TRAINS 1991 [Gross, Allen, and Traum 1993], dialogue d91-3.2).[8]

```
1.1  M: ....
1.4     first thing I'd like you to do
1.5     is send engine E2 off with a boxcar to Corning to
        pick up oranges
1.6     as soon as possible
2.1  S: okay
3.1  M: and while it's there it should pick up the tanker
```

Some of the coders in a study we carried out (Poesio and Artstein 2005) indicated the noun phrase *engine E2* as antecedent for the second *it* in utterance 3.1, whereas others indicated the immediately preceding pronoun, which they had previously marked as having *engine E2* as antecedent. Clearly, we do not want to consider these coders to be in disagreement. A solution to this dilemma has been proposed by Passonneau (2004): Use the emerging coreference sets as the 'labels' for the purpose of calculating agreement. This requires using weighted measures for calculating agreement on such sets, and

8 ftp://ftp.cs.rochester.edu/pub/papers/ai/92.tn1.trains_91_dialogues.txt.

consequently it raises serious questions about weighted measures—in particular, about the interpretability of the results, as we will see shortly.

4.4.1 Passonneau's Proposal. Passonneau (2004) recommends measuring agreement on anaphoric annotation by using *sets* of mentions of discourse entities as labels, that is, the emerging anaphoric/coreference chains. This proposal is in line with the methods developed to evaluate anaphora resolution systems (Vilain et al. 1995). But using anaphoric chains as labels would not make unweighted measures such as K a good measure for agreement. Practical experience suggests that, except when a text is very short, few annotators will catch all mentions of a discourse entity: Most will forget to mark a few, with the result that the chains (that is, category labels) differ from coder to coder and agreement as measured with K is always very low. What is needed is a coefficient that also allows for partial disagreement between judgments, when two annotators agree on part of the coreference chain but not on all of it.

Passonneau (2004) suggests solving the problem by using α with a distance metric that allows for partial agreement among anaphoric chains. Passonneau proposes a distance metric based on the following rationale: Two sets are minimally distant when they are identical and maximally distant when they are disjoint; between these extremes, sets that stand in a subset relation are closer (less distant) than ones that merely intersect. This leads to the following distance metric between two sets A and B.

$$\mathbf{d}_P = \begin{cases} 0 \text{ if } A = B \\ 1/3 \text{ if } A \subset B \text{ or } B \subset A \\ 2/3 \text{ if } A \cap B \neq \emptyset, \text{ but } A \not\subset B \text{ and } B \not\subset A \\ 1 \text{ if } A \cap B = \emptyset \end{cases}$$

Alternative distance metrics take the size of the anaphoric chain into account, based on measures used to compare sets in Information Retrieval, such as the coefficient of community of Jaccard (1912) and the coincidence index of Dice (1945) (Manning and Schütze 1999).

$$\text{Jaccard: } \mathbf{d}_J = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad \text{Dice: } \mathbf{d}_D = 1 - \frac{2\,|A \cap B|}{|A| + |B|}$$

In later work, Passonneau (2006) offers a refined distance metric which she called MASI (Measuring Agreement on Set-valued Items), obtained by multiplying Passonneau's original metric $\mathbf{d}_P$ by the metric derived from Jaccard $\mathbf{d}_J$.

$$\mathbf{d}_M = \mathbf{d}_P \times \mathbf{d}_J$$

4.4.2 Experience with α for Anaphoric Annotation. In the experiment mentioned previously (Poesio and Artstein 2005) we used 18 coders to test α and K under a variety of conditions. We found that even though our coders by and large agreed on the interpretation of anaphoric expressions, virtually no coder ever identified all the mentions of a discourse entity. As a result, even though the values of α and K obtained by using the ID of the antecedent as label were pretty similar, the values obtained when using anaphoric chains as labels were drastically different. The value of α increased, because examples where coders linked a markable to different antecedents in the same chain were no

longer considered as disagreements. However, the value of K was drastically reduced, because hardly any coder identified all the mentions of discourse entities (Figure 3).

The study also looked at the matter of individual annotator bias, and as mentioned in Section 3.1, we did not find differences between α and a κ-style version of α beyond the third decimal point. This similarity is what one would expect, given the result about annotator bias from Section 3.1 and given that in this experiment we used 18 annotators. These very small differences should be contrasted with the differences resulting from the choice of distance metrics, where values for the full-chain condition ranged from $\alpha = 0.642$ using Jaccard as distance metric, to $\alpha = 0.654$ using Passonneau's metric, to the value for Dice reported in Figure 3, $\alpha = 0.691$. These differences raise an important issue concerning the application of α-like measures for CL tasks: Using α makes it difficult to compare the results of different annotation experiments, in that a "poor" value or a "high" value might result from "too strict" or "too generous" distance metrics, making it even more important to develop a methodology to identify appropriate values for these coefficients. This issue is further emphasized by the study reported next.

4.4.3 Discourse Deixis. A second annotation study we carried out (Artstein and Poesio 2006) shows even more clearly the possible side effects of using weighted coefficients. This study was concerned with the annotation of the antecedents of references to abstract objects, such as the example of the pronoun *that* in utterance 7.6 (TRAINS 1991, dialogue d91-2.2).

```
7.3  : so we ship one
7.4  : boxcar
7.5  : of oranges to Elmira
7.6  : and that takes another 2 hours
```

Previous studies of discourse deixis annotation showed that these are extremely difficult judgments to make (Eckert and Strube 2000; Navarretta 2000; Byron 2002), except perhaps for identifying the type of object (Poesio and Modjeska 2005), so we simplified the task by only requiring our participants to identify the boundaries of the area of text in which the antecedent was introduced. Even so, we found a great variety in how these boundaries were marked: Exactly as in the case of discourse segmentation discussed earlier, our participants broadly agreed on the area of text, but disagreed on

Chain	K	α
None	0.628	0.656
Partial	0.563	0.677
Full	0.480	0.691

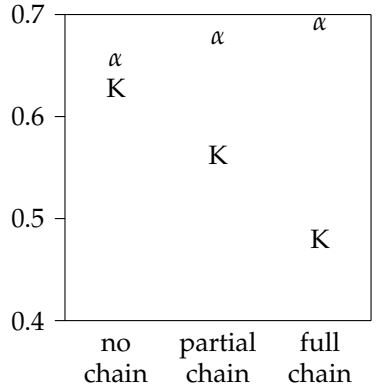

Figure 3
A comparison of the values of α and K for anaphoric annotation (Poesio and Artstein 2005).

its exact boundary. For instance, in this example, nine out of ten annotators marked the antecedent of *that* as a text segment ending with the word *Elmira*, but some started with the word *so*, some started with *we*, some with *ship*, and some with *one*.

We tested a number of ways to measure partial agreement on this task, and obtained widely different results. First of all, we tested three set-based distance metrics inspired by the Passonneau proposals that we just discussed: We considered discourse segments to be sets of words, and computed the distance between them using Passonneau's metric, Jaccard, and Dice. Using these three metrics, we obtained α values of 0.55 (with Passonneau's metric), 0.45 (with Jaccard), and 0.55 (with Dice). We should note that because antecedents of different expressions rarely overlapped, the expected disagreement was close to 1 (maximal), so the value of α turned out to be very close to the complement of the observed disagreement as calculated by the different distance metrics.

Next, we considered methods based on the position of words in the text. The first method computed differences between absolute boundary positions: Each antecedent was associated with the position of its first or last word in the dialogue, and agreement was calculated using α with the interval distance metric. This gave us α values of 0.998 for the beginnings of the antecedent-evoking area and 0.999 for the ends. This is because expected disagreement is exceptionally low: Coders tend to mark discourse antecedents close to the referring expression, so the average distance between antecedents of the same expression is smaller than the size of the dialogue by a few orders of magnitude. The second method associated each antecedent with the position of its first or last word *relative to the beginning of the anaphoric expression*. This time we found extremely low values of $\alpha = 0.167$ for beginnings of antecedents and 0.122 for ends—barely in the positive side. This shows that agreement among coders is not dramatically better than what would be expected if they just marked discourse antecedents at a fixed distance from the referring expression.

The three ranges of α that we observed (middle, high, and low) show agreement on the identity of discourse antecedents, their position in the dialogue, and their position relative to referring expressions, respectively. The middle range shows variability of up to 10 percentage points, depending on the distance metric chosen. The lesson is that once we start using weighted measures we cannot anymore interpret the value of α using traditional rules of thumb such as those proposed by Krippendorff or by Landis and Koch. This is because depending on the way we measure agreement, we can report α values ranging from 0.122 to 0.998 for the very same experiment! New interpretation methods have to be developed, which will be task- and distance-metric specific. We'll return to this issue in the conclusions.

4.5 Word Senses

Word sense tagging is one of the hardest annotation tasks. Whereas in the case of part-of-speech and dialogue act tagging the same categories are used to classify all units, in the case of word sense tagging different categories must be used for each word, which makes writing a single coding manual specifying examples for all categories impossible: The only option is to rely on a dictionary. Unfortunately, different dictionaries make different distinctions, and often coders can't make the fine-grained distinctions that trained lexicographers can make. The problem is particularly serious for verbs, which tend to be polysemous rather than homonymous (Palmer, Dang, and Fellbaum 2007).

These difficulties, and in particular the difficulty of tagging senses with a fine-grained repertoire of senses such as that provided by dictionaries or by WordNet (Fellbaum 1998), have been highlighted by the three SENSEVAL initiatives. Already

during the first SENSEVAL, Véronis (1998) carried out two studies of intercoder agreement on word sense tagging in the so-called ROMANSEVAL task. One study was concerned with agreement on polysemy—that is, the extent to which coders agreed that a word was polysemous in a given context. Six naive coders were asked to make this judgment about 600 French words (200 nouns, 200 verbs, 200 adjectives) using the repertoire of senses in the *Petit Larousse*. On this task, a (pairwise) percentage agreement of 0.68 for nouns, 0.74 for verbs, and 0.78 for adjectives was observed, corresponding to K values of 0.36, 0.37, and 0.67, respectively. The 20 words from each category perceived by the coders in this first experiment to be most polysemous were then used in a second study, of intercoder agreement on the sense tagging task, which involved six different naive coders. Interestingly, the coders in this second experiment were allowed to assign multiple tags to words, although they did not make much use of this possibility; so κ_w was used to measure agreement. In this experiment, Véronis observed (weighted) pairwise agreement of 0.63 for verbs, 0.71 for adjectives, and 0.73 for nouns, corresponding to κ_w values of 0.41, 0.41, and 0.46, but with a wide variety of values when measured per word—ranging from 0.007 for the adjective *correct* to 0.92 for the noun *détention*. Similarly mediocre results for intercoder agreement between naive coders were reported in the subsequent editions of SENSEVAL. Agreement studies for SENSEVAL-2, where WordNet senses were used as tags, reported a percentage agreement for verb senses of around 70%, whereas for SENSEVAL-3 (English Lexical Sample Task), Mihalcea, Chklovski, and Kilgarriff (2004) report a percentage agreement of 67.3% and average K of 0.58.

Two types of solutions have been proposed for the problem of low agreement on sense tagging. The solution proposed by Kilgarriff (1999) is to use professional lexicographers and arbitration. The study carried out by Kilgarriff does not therefore qualify as a true study of replicability in the sense of the terms used by Krippendorff, but it did show that this approach makes it possible to achieve percentage agreement of around 95.5%. An alternative approach has been to address the problem of the inability of naive coders to make fine-grained distinctions by introducing coarser-grained classification schemes which group together dictionary senses (Bruce and Wiebe, 1998; Buitelaar 1998; Véronis 1998; Palmer, Dang, and Fellbaum 2007). Hierarchical tagsets were also developed, such as HECTOR (Atkins 1992) or, indeed, WordNet itself (where senses are related by hyponymy links). In the case of Buitelaar and Palmer, Dang, and Fellbaum, the "supersenses" were identified by hand, whereas Bruce and Wiebe and Véronis used clustering methods such as those from Bruce and Wiebe (1999) to collapse some of the initial sense distinctions.[9] Palmer, Dang, and Fellbaum (2007) illustrate this practice with the example of the verb *call*, which has 28 fine-grained senses in WordNet 1.7: They conflate these senses into a small number of groups using various criteria—for example, four senses can be grouped in a group they call Group 1 on the basis of subcategorization frame similarities (Table 9).

Palmer, Dang, and Fellbaum (2007) achieved for the English Verb Lexical Sense task of SENSEVAL-2 a percentage agreement among coders of 82% with grouped senses, as opposed to 71% with the original WordNet senses. Bruce and Wiebe (1998) found that collapsing the senses of their test word (*interest*) on the basis of their use by coders and merging the two classes found to be harder to distinguish resulted in an increase of

9 The methodology proposed in Bruce and Wiebe (1999) is in our view the most advanced technique to "make sense" of the results of agreement studies available in the literature. The extended version of this article contains a fuller introduction to these methods.

Table 9
Group 1 of senses of *call* in Palmer, Dang, and Fellbaum (2007, page 149).

SENSE	DESCRIPTION	EXAMPLE	HYPERNYM
WN1	name, call	"They named[a] their son David"	**LABEL**
WN3	call, give a quality	"She called her children lazy and ungrateful"	**LABEL**
WN19	call, consider	"I would not call her beautiful"	**SEE**
WN22	address, call	"Call me mister"	**ADDRESS**

[a] The verb *named* appears in the original WordNet example for the verb *call*.

the value of K from 0.874 to 0.898. Using a related technique, Véronis (1998) found that agreement on noun word sense tagging went up from a K of around 0.45 to a K of 0.86. We should note, however, that the post hoc merging of categories is not equivalent to running a study with fewer categories to begin with.

Attempts were also made to develop techniques to measure partial agreement with hierarchical tagsets. A first proposal in this direction was advanced by Melamed and Resnik (2000), who developed a coefficient for hierarchical tagsets that could be used in SENSEVAL for measuring agreement with tagsets such as HECTOR. Melamed and Resnik proposed to "normalize" the computation of observed and expected agreement by taking each label which is not a leaf in the tag hierarchy and distributing it down to the leaves in a uniform way, and then only computing agreement on the leaves. For example, with a tagset like the one in Table 9, the cases in which the coders used the label 'Group 1' would be uniformly "distributed down" and added in equal measure to the number of cases in which the coders assigned each of the four WordNet labels. The method proposed in the paper has, however, problematic properties when used to measure intercoder agreement. For example, suppose tag A dominates two sub-tags A1 and A2, and that two coders mark a particular item as A. Intuitively, we would want to consider this a case of perfect agreement, but this is not what the method proposed by Melamed and Resnik yields. The annotators' marks are distributed over the two sub-tags, each with probability 0.5, and then the agreement is computed by summing the joint probabilities over the two subtags (Equation (4) of Melamed and Resnik 2000), with the result that the agreement over the item turns out to be $0.5^2 + 0.5^2 = 0.5$ instead of 1. To correct this, Dan Melamed (personal communication) suggested replacing the product in Equation (4) with a minimum operator. However, the calculation of expected agreement (Equation (5) of Melamed and Resnik 2000) still gives the amount of agreement which is expected if coders are forced to choose among leaf nodes, which makes this method inappropriate for coding schemes that do not force coders to do this.

One way to use Melamed and Resnik's proposal while avoiding the discrepancy between observed and expected agreement is to treat the proposal not as a new coefficient, but rather as a distance metric to be plugged into a weighted coefficient like α. Let A and B be two nodes in a hierarchical tagset, let L be the set of all leaf nodes in the tagset, and let $\mathrm{P}(l|T)$ be the probability of selecting a leaf node l given an arbitrary node T when the probability mass of T is distributed uniformly to all the nodes dominated by T. We can reinterpret Melamed's modification of Equation (4) in Melamed and Resnik (2000) as a metric measuring the distance between nodes A and B.

$$\mathbf{d}_{M+R} = 1 - \sum_{l \in L} \min(\mathrm{P}(l|A), \mathrm{P}(l|B))$$

This metric has the desirable properties—it is 0 when tags A and B are identical, 1 when the tags do not overlap, and somewhere in between in all other cases. If we use this metric for Krippendorff's α we find that observed agreement is exactly the same as in Melamed and Resnik (2000) with the product operator replaced by minimum (Melamed's modification).

We can also use other distance metrics with α. For example, we could associate with each sense an **extended sense**—a set **es**(s) including the sense itself and its grouped sense—and then use set-based distance metrics from Section 4.4, for example Passonneau's $\mathbf{d}_P$. To illustrate how this approach could be used to measure (dis)agreement on word sense annotation, suppose that two coders have to annotate the use of *call* in the following sentence (from the WSJ part of the Penn Treebank, section 02, text w0209):

> This gene, **called** "gametocide," is carried into the plant by a virus that remains active for a few days.

The standard guidelines (in SENSEVAL, say) require coders to assign a WN sense to words. Under such guidelines, if coder A classifies the use of *called* in the above example as an instance of WN1, whereas coder B annotates it as an instance of WN3, we would find total disagreement ($\mathbf{d}_{k_a k_b} = 1$) which seems excessively harsh as the two senses are clearly related. However, by using the broader senses proposed by Palmer, Dang, and Fellbaum (2007) in combination with a distance metric such as the one just proposed, it is possible to get more flexible and, we believe, more realistic assessments of the degree of agreement in situations such as this. For instance, in case the reliability study had already been carried out under the standard SENSEVAL guidelines, the distance metric proposed above could be used to identify post hoc cases of partial agreement by adding to each WN sense its hypernyms according to the groupings proposed by Palmer, Dang, and Fellbaum. For example, A's annotation could be turned into a new set label {WN1,**LABEL**} and B's mark into the set table {WN3,**LABEL**}, which would give a distance $\mathbf{d} = 2/3$, indicating a degree of overlap. The method for computing agreement proposed here could could also be used to allow coders to choose either a more specific label or one of Palmer, Dang, and Fellbaum's superlabels. For example, suppose A sticks to WN1, but B decides to mark the use above using Palmer, Dang, and Fellbaum's **LABEL** category, then we would still find a distance $\mathbf{d} = 1/3$.

An alternative way of using α for word sense annotation was developed and tested by Passonneau, Habash, and Rambow (2006). Their approach is to allow coders to assign multiple labels (WordNet synsets) for wordsenses, as done by Véronis (1998) and more recently by Rosenberg and Binkowski (2004) for text classification labels and by Poesio and Artstein (2005) for anaphora. These multi-label sets can then be compared using the MASI distance metric for α (Passonneau 2006).

5. Conclusions

The purpose of this article has been to expose the reader to the mathematics of chance-corrected coefficients of agreement as well as the current state of the art of using these coefficients in CL. Our hope is that readers come to view agreement studies not as an additional chore or hurdle for publication, but as a tool for analysis which offers new insights into the annotation process. We conclude by summarizing what in our view are the main recommendations emerging from ten years of experience with coefficients of agreement. These can be grouped under three main headings: methodology, choice of coefficients, and interpretation of coefficients.

5.1 Methodology

Our first recommendation is that annotation efforts should perform and report rigorous reliability testing. The last decade has already seen considerable improvement, from the absence of any tests for the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993) or the British National Corpus (Leech, Garside, and Bryant 1994) to the central role played by reliability testing in the Penn Discourse Treebank (Miltsakaki et al. 2004) and OntoNotes (Hovy et al. 2006). But even the latter efforts only measure and report percent agreement. We believe that part of the reluctance to report chance-corrected measures is the difficulty in interpreting them. However, our experience is that chance-corrected coefficients of agreement do provide a better indication of the quality of the resulting annotation than simple percent agreement, and moreover, the detailed calculations leading to the coefficients can be very revealing as to where the disagreements are located and what their sources may be.

A rigorous methodology for reliability testing does not, in our opinion, exclude the use of expert coders, and here we feel there may be a motivated difference between the fields of content analysis and CL. There is a clear tradeoff between the complexity of the judgments that coders are required to make and the reliability of such judgments, and we should strive to devise annotation schemes that are not only reliable enough to be replicated, but also sophisticated enough to be useful (cf. Krippendorff 2004a, pages 213–214). In content analysis, conclusions are drawn directly from annotated corpora, so the emphasis is more on replicability; whereas in CL, corpora constitute a resource which is used by other processes, so the emphasis is more towards usefulness. There is also a tradeoff between the sophistication of judgments and the availability of coders who can make such judgments. Consequently, annotation by experts is often the only practical way to get useful corpora for CL. Current practice achieves high reliability either by using professionals (Kilgarriff 1999) or through intensive training (Hovy et al. 2006; Carlson, Marcu, and Okurowski 2003); this means that results are not replicable across sites, and are therefore less reliable than annotation by naive coders adhering to written instructions. We feel that inter-annotator agreement studies should still be carried out, as they serve as an assurance that the results are replicable when the annotators are chosen from the same population as the original annotators. An important additional assurance should be provided in the form of an independent evaluation of the task for which the corpus is used (cf. Passonneau 2006).

5.2 Choosing a Coefficient

One of the goals of this article is to help authors make an informed choice regarding the coefficients they use for measuring agreement. While coefficients other than K, specifically Cohen's κ and Krippendorff's α, have appeared in the CL literature as early as Carletta (1996) and Passonneau and Litman (1996), they hadn't sprung into general awareness until the publication of Di Eugenio and Glass (2004) and Passonneau (2004). Regarding the question of annotator bias, there is an overwhelming consensus in CL practice: K and α are used in the vast majority of the studies we reported. We agree with the view that K and α are more appropriate, as they abstract away from the bias of specific coders. But we also believe that ultimately this issue of annotator bias is of little consequence because the differences get smaller and smaller as the number of annotators grows (Artstein and Poesio 2005). We believe that increasing the number of annotators is the best strategy, because it reduces the chances of accidental personal biases.

However, Krippendorff's α is indispensable when the category labels are not equally distinct from one another. We think there are at least two types of coding schemes in which this is the case: (i) hierarchical tagsets and (ii) set-valued interpretations such as those proposed for anaphora. At least in the second case, weighted coefficients are almost unavoidable. We therefore recommend using α, noting however that the specific choice of weights will affect the overall numerical result.

5.3 Interpreting the Values

We view the lack of consensus on how to interpret the values of agreement coefficients as a serious problem with current practice in reliability testing, and as one of the main reasons for the reluctance of many in CL to embark on reliability studies. Unlike significance values which report a probability (that an observed effect is due to chance), agreement coefficients report a magnitude, and it is less clear how to interpret such magnitudes. Our own experience is consistent with that of Krippendorff: Both in our earlier work (Poesio and Vieira 1998; Poesio 2004a) and in the more recent efforts (Poesio and Artstein 2005) we found that only values above 0.8 ensured an annotation of reasonable quality (Poesio 2004a). We therefore feel that if a threshold needs to be set, 0.8 is a good value.

That said, we doubt that a single cutoff point is appropriate for all purposes. For some CL studies, particularly on discourse, useful corpora have been obtained while attaining reliability only at the 0.7 level. We agree therefore with Craggs and McGee Wood (2005) that setting a specific agreement threshold should not be a prerequisite for publication. Instead, as recommended by Di Eugenio and Glass (2004) and others, researchers should report in detail on the methodology that was followed in collecting the reliability data (number of coders, whether they coded independently, whether they relied exclusively on an annotation manual), whether agreement was statistically significant, and provide a confusion matrix or agreement table so that readers can find out whether overall figures of agreement hide disagreements on less common categories. For an example of good practice in this respect, see Teufel and Moens (2002). The decision whether a corpus is good enough for publication should be based on more than the agreement score—specifically, an important consideration is an independent evaluation of the results that are based on the corpus.

Acknowledgments

This work was supported in part by EPSRC grant GR/S76434/01, ARRAU. We wish to thank four anonymous reviewers and Jean Carletta, Mark Core, Barbara Di Eugenio, Ruth Filik, Michael Glass, George Hripcsak, Adam Kilgarriff, Dan Melamed, Becky Passonneau, Phil Resnik, Tony Sanford, Patrick Sturt, and David Traum for helpful comments and discussion. Special thanks to Klaus Krippendorff for an extremely detailed review of an earlier version of this article. We are also extremely grateful to the British Library in London, which made accessible to us virtually every paper we needed for this research.

References

Allen, James and Mark Core. 1997. DAMSL: Dialogue act markup in several layers. Draft contribution for the Discourse Resource Initiative, University of Rochester. Available at `http://www.cs.rochester.edu/research/cisd/resources/damsl/`.

Artstein, Ron and Massimo Poesio. 2005. Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL 2005*, pages 141–150, Edinburgh.

Artstein, Ron and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *brandial 2006: Proceedings of the 10th Workshop on the Semantics and*

Pragmatics of Dialogue, pages 56–63, Potsdam.
Atkins, Sue. 1992. Tools for computer-aided corpus lexicography: The Hector project. *Acta Linguistica Hungarica*, 41:5–71.
Babarczy, Anna, John Carroll, and Geoffrey Sampson. 2006. Definitional, personal, and mechanical constraints on part of speech annotation performance. *Natural Language Engineering*, 12(1):77–90.
Bartko, John J. and William T. Carpenter, Jr. 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163(5):307–317.
Beeferman, Doug, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.
Bennett, E. M., R. Alpert, and A. C. Goldstein. 1954. Communications through limited questioning. *Public Opinion Quarterly*, 18(3):303–308.
Bloch, Daniel A. and Helena Chmura Kraemer. 1989. 2×2 kappa coefficients: Measures of agreement or association. *Biometrics*, 45(1):269–287.
Brennan, Robert L. and Dale J. Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687–699.
Bruce, Rebecca and Janyce Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of EMNLP*, pages 53–60, Granada.
Bruce, Rebecca F. and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.
Buitelaar, Paul. 1998. *CoreLex : Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University, Waltham, MA.
Bunt, Harry C. 2000. Dynamic interpretation and dialogue theory. In Martin M. Taylor, Françoise Néel, and Don G. Bouwhuis, editors, *The Structure of Multimodal Dialogue II*. John Benjamins, Amsterdam, pages 139–166.
Bunt, Harry C. 2005. A framework for dialogue act specification. In *Proceedings of the Joint ISO-ACL Workshop on the Representation and Annotation of Semantic Information*, Tilburg. Available at: `http://let.uvt.nl/research/ti/sigsem/wg/discussionnotes4.htm`.
Byron, Donna K. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 80–87, Philadelphia, PA.
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):13–32.
Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan C. J. van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*. Kluwer, Dordrecht, pages 85–112.
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(1):37–46.
Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
Core, Mark G. and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, AAAI, Cambridge, MA. Available at: `http://www.cs.umd.edu/~traum/CA/fpapers.html`.
Craggs, Richard and Mary McGee Wood. 2004. A two-dimensional annotation scheme for emotion in dialogue. In *Papers from the 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, pages 44–49.
Craggs, Richard and Mary McGee Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–295.
Davies, Mark and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.
Di Eugenio, Barbara. 2000. On the usage of Kappa to evaluate agreement on coding tasks. In *Proceedings of LREC*, volume 1, pages 441–444, Athens.
Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Di Eugenio, Barbara, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An empirical investigation of proposals in collaborative dialogues. In *Proceedings of 36th Annual Meeting of the ACL*, pages 325–329, Montreal.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Donner, Allan and Michael Eliasziw. 1987. Sample size requirements for reliability studies. *Statistics in Medicine*, 6:441–448.

Doran, Christine, John Aberdeen, Laurie Damianos, and Lynette Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark. Available at: `http://www.sigdial.org/workshops/workshop2/proceedings`.

Eckert, Miriam and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fleiss, Joseph L. 1975. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31(3):651–659.

Francis, W. Nelson and Henry Kucera. 1982. *Frequency Analysis of English Usage: lexicon and grammar*. Houghton Mifflin, Boston, MA.

Geertzen, Jeroen and Harry Bunt. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 126–133, Sydney.

Gross, Derek, James F. Allen, and David R. Traum. 1993. The Trains 91 dialogues. TRAINS Technical Note 92-1, University of Rochester Computer Science Department, Rochester, NY.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hayes, Andrew F. and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT–NAACL, Companion Volume: Short Papers*, pages 57–60, New York.

Hsu, Louis M. and Ronald Field. 2003. Interrater agreement measures: Comments on kappa$_n$, Cohen's kappa, Scott's π, and Aickin's α. *Understanding Statistics*, 2(3):205–219.

Jaccard, Paul. 1912. The distribution of the flora in the Alpine zone. *New Phytologist*, 11(2):37–50.

Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. 1995. Dialogue acts in VERBMOBIL. VM-Report 65, Universität Hamburg, DFKI GmbH, Universität Erlangen, and TU Berlin.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado at Boulder, Institute for Cognitive Science.

Kilgarriff, Adam. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 277–278, Bergen, Norway.

Klavans, Judith L., Samuel Popper, and Rebecca Passonneau. 2003. Tackling the internet glossary glut: Automatic extraction and evaluation of genus phrases. In *Proceedings of the SIGIR-2003 Workshop on the Semantic Web*, Toronto.

Kowtko, Jacqueline C., Stephen D. Isard, and Gwyneth M. Doherty. 1992. Conversational games within dialogue. Research Paper HCRC/RP-31, Human Communication Research Centre, University of Edinburgh.

Krippendorff, Klaus. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Krippendorff, Klaus. 1978. Reliability of binary attribute data. *Biometrics*, 34(1):142–144. Letter to the editor, with a reply by Joseph L. Fleiss.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA.

Krippendorff, Klaus. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, 25:47–76.

Krippendorff, Klaus. 2004a. *Content Analysis: An Introduction to Its Methodology*, second edition, chapter 11. Sage, Thousand Oaks, CA.

Krippendorff, Klaus. 2004b. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of COLING 1994: The 15th International Conference on Computational Linguistics, Volume 1*, pages 622–628, Kyoto.

Levin, James A. and James A. Moore. 1978. Dialogue-games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420.

Manning, Christopher D. and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcu, Daniel, Magdalena Romera, and Estibaliz Amorrortu. 1999. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Workshop on Levels of Representation in Discourse*, pages 71–78, University of Edinburgh.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marion, Rodger. 2004. The whole art of deduction. Unpublished manuscript.

Melamed, I. Dan and Philip Resnik. 2000. Tagger evaluation given hierarchical tagsets. *Computers and the Humanities*, 34(1–2):79–84. Available at: `http://www.sahs/utmb.edu/PELLINORE/Intro_to_research/wad/wad/_home.htm`.

Mieskes, Margot and Michael Strube. 2006. Part-of-speech tagging of transcribed speech. In *Proceedings of LREC*, pages 935–938, Genoa.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The SENSEVAL-3 English lexical sample task. In *Proceedings of SENSEVAL-3*, pages 25–28, Barcelona.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, MA.

Moser, Megan G., Johanna D. Moore, and Erin Glendening. 1996. Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units. Technical Report 96-17, University of Pittsburgh, Department of Computer Science.

Navarretta, Costanza. 2000. Abstract anaphora resolution in Danish. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue*, Hong Kong, pages 56–65.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL 2004*, pages 145–152, Boston, MA.

Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA.

Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Passonneau, Rebecca J. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*, volume 4, pages 1503–1506, Lisbon.

Passonneau, Rebecca J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of LREC*, Genoa, pages 831–836.

Passonneau, Rebecca J., Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of LREC*, Genoa, pages 1951–1956.

Passonneau, Rebecca J. and Diane J. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of 31st Annual Meeting of the ACL*, pages 148–155, Columbus, OH.

Passonneau, Rebecca J. and Diane J. Litman. 1996. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence and linguistic devices. In Eduard H. Hovy and Donia R. Scott, editors, *Computational and Conversational Discourse: Burning Issues – An Interdisciplinary Account*, volume 151 of *NATO ASI Series F: Computer and Systems Sciences*. Springer, Berlin, chapter 7, pages 161–194.

Passonneau, Rebecca J. and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Pevzner, Lev and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Poesio, Massimo. 2004a. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona.

Poesio, Massimo. 2004b. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, MA.

Poesio, Massimo and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, MI.

Poesio, Massimo and Natalia N. Modjeska. 2005. Focus, activation, and *this*-noun phrases: An empirical study. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing*, volume 263 of *Current Issues in Linguistic Theory*. John Benjamins, pages 429–442, Amsterdam and Philadelphia.

Poesio, Massimo, A. Patel, and Barbara Di Eugenio. 2006. Discourse structure and anaphora in tutorial dialogues: An empirical analysis of two theories of the global focus. *Research in Language and Computation*, 4(2–3):229–257.

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Popescu-Belis, Andrei. 2005. Dialogue acts: One or more dimensions? Working Paper 62, ISSCO, University of Geneva.

Posner, Karen L., Paul D. Sampson, Robert A. Caplan, Richard J. Ward, and Frederick W. Cheney. 1990. Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine*, 9:1103–1115.

Rajaratnam, Nageswari. 1960. Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, 25(3):261–271.

Reidsma, Dennis and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Reinhart, T. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1):53–93.

Reynar, Jeffrey C. 1998. *Topic Segmentation: Algorithms and Applications*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Ries, Klaus. 2002. Segmenting conversations by topic, initiative and style. In Anni R. Coden, Eric W. Brown, and Savitha Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Lecture Notes in Computer Science*. Springer, Berlin, pages 51–66.

Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin.

Rosenberg, Andrew and Ed Binkowski. 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 77–80, Boston, MA.

Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, MA.

Siegel, Sidney and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition, chapter 9.8. McGraw-Hill, New York.

Stent, Amanda J. 2001. *Dialogue Systems as Conversational Partners: Applying Conversation Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue*. Ph.D. thesis, Department of Computer Science, University of Rochester.

Stevenson, Mark and Robert Gaizauskas. 2000. Experiments on sentence boundary detection. In *Proceedings of 6th ANLP*, pages 84–89, Seattle, WA.

Stolcke, Andreas, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Stuart, Alan. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3/4):412–416.

Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of Ninth Conference of the EACL*, pages 110–117, Bergen.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Traum, David R. and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599.

Vallduví, Enric. 1993. Information packaging: A survey. Research Paper RP-44, University of Edinburgh, HCRC.

Véronis, Jean. 1998. A study of polysemy judgments and inter-annotator agreement. In *Proceedings of SENSEVAL-1*, Herstmonceux Castle, England. Available at: `http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/PROCEEDINGS/`.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference*, pages 45–52, Columbia, MD.

Zwick, Rebecca. 1988. Another look at interrater agreement. *Psychological Bulletin*, 103(3):374–378.