Survey Article

A shortened version of this article was submitted to the journal Computational Linguistics; this is the full version.

Inter-Coder Agreement for Computational Linguistics

Ron Artstein
University of Essex

Massimo Poesio
Università di Trento / University of Essex

This article is a survey of issues concerning the measurement of agreement among corpus annotators. It exposes the mathematics and underlying assumptions of agreement coefficients such as Cohen's kappa and Krippendorff's alpha; relates these coefficients to explicit models of annotator error; 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

Ever since the mid-Nineties, 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 to the first resource creation efforts such as the creation of the Brown corpus (Francis and Kucera 1982), the British National Corpus (Leech, Garside, and Bryant 1994) or the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 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 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; Stolcke et al. 1997; Poesio and Vieira 1998; Di Eugenio 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; Nenkova and Passonneau 2004; Mieskes and Strube 2006). During this period, however, a number of questions have

1 As we will see below, there are lots of terminological inconsistencies in the literature. Carletta uses the term kappa for the coefficient of agreement, referring to Krippendorff (1980) and Siegel and Castellan (1988) for an introduction, 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 term κ to indicate the parameter which is estimated by K (i.e. a function of K with an approximately normal distribution which can be used to estimate the significance of the value of K obtained). In what follows, we will use the term κ to indicate coefficients that calculate chance agreement by looking at individual coder marginals – Cohen's original coefficient and its generalization to more than two coders – and use the term K for the coefficient discussed by Siegel and Castellan.

also been raised about K and similar coefficients – some already in Carletta et al.'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 (2000) raised the issues of the effect of **skewed distributions** on the value of K and pointed out that the original κ developed by Cohen (1960) is based on very different assumptions about coder bias from 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 (2004) 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 dreaded 0.67 threshold, and recommended reporting both values. Craggs and McGee Wood (2005), by contrast, 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; Passonneau, Habash, and Rambow 2006; Nenkova and Passonneau 2004).

Now that more than ten years have passed since Carletta's original presentation at the workshop on Empirical Methods in Discourse, we feel it is the 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 of all, we discuss in some detail the mathematics and underlying assumptions of the coefficients used or mentioned in the CL or the content analysis literature. Secondly, we also cover in some detail Krippendorff's α, often mentioned but never really discussed in detail in previous CL literature other than 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. Coefficients of Agreement

2.1 Agreement, reliability and validity

To begin with, a quick recap of what agreement studies can and cannot achieve. The following section is inspired by Krippendorff (2004a, Section 11.1).

Researchers that 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**. The fundamental assumption behind the methodologies discussed in this paper 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.

The last point to keep in mind is that the term 'reliability' can be used in three different ways, depending on how agreement is tested. First of all, we may want to test **stability**, or **intra-coder agreement**: the extent to which the coding process yields the same results when repeated over time, typically measured by observing how much the same coder agrees with her or his previous coding at a distance of time. Measuring stability is often the first step towards assessing the reliability of data. A stronger test is measuring **reproducibility**: the degree to which different coders achieve the same coding when working independently. This is the type of test required for large annotation efforts employing multiple coders. Finally, **accuracy** is the degree to which a coding process yields the results specified by a gold standard, when one such exists.

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 in the literature often use different notations to express these concepts. We will 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 notations uniformly throughout the article.

- We will use the notation A_o (observed agreement) and D_o (observed disagreement) to indicate the observed agreement and disagreement.
- The notation A_e and D_e will be used to indicate 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).
- The notation $P(\cdot)$ will be reserved for the probability of a variable, and $\hat{P}(\cdot)$ will be used for an estimate of such probability from observed data.

Finally, we will use **n** with a subscript parameter to indicate the number of judgments of a particular 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;

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

- $\mathbf{n}_k$ is the total number of items assigned by all coders to category k.

2.3 Unsatisfactory Measures of Agreement

Why are new coefficients to measure agreement necessary? Couldn't existing measures such as percentage agreement or traditional statistics like χ^2 do the job? Although this question has already been addressed a number of times in the literature, it is useful to consider it again, in part for completeness' sake, but also to clarify the problems that kappa-like measures are meant to solve.

Percentage Agreement. 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 discussion below, 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:

$$\text{A}_\text{o} = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i$$

For example, let us assume we have a very simple annotation scheme for dialogue acts in information-seeking dialogues making a binary distinction between the categories `statement` and `info-request`, as in the DAMSL dialogue act scheme (Allen and Core 1997), and that 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: $\text{A}_\text{o} = (20 + 50)/100 = 0.7$

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, since 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 num-

ber 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 ($\frac{1}{9} + \frac{1}{9} + \frac{1}{9}$). The second reason percentage agreement can not be trusted is that is 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. 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. We will review various ways of correcting percentage agreement for chance, starting in section 2.4. But first, we will look at other chance-adjusted measures and show why they are inadequate for measuring agreement – hence the need to develop specific agreement coefficients.

Measures of association. The χ^2 statistic is also inappropriate as a measure of agreement. As pointed out by Cohen (1960, page 39), χ^2 is a measure of association rather than agreement – which means that we get a high value of χ^2 whenever a particular cooccurrence of judgments is different from the expected value. This may happen not just when we find good agreement, but also when we have systematic disagreement. The agreement matrix in Table 2 reports the results of an annotation experiment in which again coder A and coder B classify utterances as either `statement`, `info-request`, or `check`. The value of χ^2 for this table is 64.59, which is highly significant. But this strong association does not indicate agreement: the highest contribution comes from the utterances classified by A as `info-request` and by B as `check`, where the observed value 0.15 is much higher than the expected value 0.06 – a case of disagreement.

Correlation Coefficients. A point perhaps not sufficiently emphasized in the CL literature on agreement is that κ and related measures of agreement such as α or π are not primarily statistics in the sense of t, χ^2 or F, which are (functions associated with) probability distributions whose value specifies the significance of the result obtained. The title of Cohen's classic article is very illuminating in this respect: π, κ, α, etc. are 'coefficient(s) of agreement for nominal scales'. What this means is that they are coefficients taking values between -1 and $+1$, just like Pearson's product-moment

Table 2
High association but low agreement (adapted from Cohen 1960).

		CODER A			
		STAT	IREQ	CHCK	TOTAL
CODER B	STAT	0.25	0.13	0.12	0.50
	IREQ	0.12	0.02	0.16	0.30
	CHCK	0.03	0.15	0.02	0.20
	TOTAL	0.40	0.30	0.30	1

Table 3
Correlation need not indicate agreement.

ITEM	EXP 1		EXP 2	
	A	B	C	D
a	1	1	1	2
b	2	2	2	4
c	3	3	3	6
d	4	4	4	8
e	5	5	5	10
	$r = 1.0$		$r = 1.0$	

coefficient r or Spearman's rank-correlation coefficient r_s, but intended for nominal scales, and for measuring agreement rather than association. Thinking of the kappa-like measures of agreement as coefficients is illuminating in certain respects, as they have some of the formal properties of correlation coefficients (Krippendorff 1970a), and the problem of deciding whether a particular value of, say, κ indicates a sufficient degree of agreement is similar to the problem of determining whether a particular value of r expresses a strong enough association. However, neither product-moment correlation r nor rank order correlation r_s are good measures of agreement (Bartko and Carpenter 1976, page 309). This is not just because these coefficients are specified for real values rather than nominal scales; correlation is not the same thing as agreement, and a strong correlation may exist even when coders disagree. The problem is illustrated by Table 3 (adapted from Bartko and Carpenter 1976). Suppose we have a coding scheme according to which coders give each item a rating between 1 and 10 (this might be a marking scheme for student essays, for example), and we ran two experiments to test the scheme. In the first experiment, coders A and B (whose marks are shown in the second and third columns) are in complete agreement; while in the second, coders C and D (whose marks are shown in the fourth and fifth columns) disagree on all items, but assign marks that are linearly correlated. Exactly the same product-moment value will be obtained in both experiments, even though there is perfect agreement between A and B, but no agreement at all between C and D.

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; whereas 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, since 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 is the product of the chance of each of them assigning an item to that category: $P(k|c_1) \cdot P(k|c_2)$ (the independence assumption has been the subject of much criticism, for example by John S. Uebersax).[2] Expected agreement is then the probability of c_1 and c_2 agreeing on any category, that is, the sum of the above product over all categories:

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

The difference between S, π and κ lies in the assumptions leading to the calculation of $P(k|c_i)$, 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, $P(k_j|c_m) = P(k_l|c_n)$. (Put it another way: chance does not distinguish between categories and coders.)
- 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,

2 `http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm`.

$P(k|c_m) = P(k|c_n)$. (i.e., chance distinguishes between categories, but not coders.)

κ: If coders were operating by chance alone, we would get a separate distribution for each coder: chance distinguishes between both categories and coders.

A further complication, explained perhaps most clearly by Krippendorff (1980), is a problem that is all too familiar in CL: the lack of independent prior knowledge of the distribution of items among categories. The distribution of categories (for π) and the priors for the individual coders (for κ) therefore have to be estimated from the observed data. We begin here by giving detailed examples on how the coefficients are calculated for two coders; we will discuss a variety of proposed generalizations starting in section 2.5.

All categories are equally likely: S. The simplest way of discounting for chance is the one adopted to compute the coefficient S, also known in the literature as C, κ_n, G, and RE (see Zwick 1988; Hsu and Field 2003). As said above, 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 $\mathbf{k}$ categories, then the chance $P(k|c_i)$ of any of 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

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

For example, the value of S for the coding example in Table 1 is as follows (where $A_o = 0.7$, see above).

$$A_e^S = 2 \times \left(\frac{1}{2}\right)^2 = 0.5$$

$$S = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

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. A uniform distribution is not a very plausible model for annotation in CL, as in pretty much all tagging tasks, from POS tags (Francis and Kucera 1982; Mieskes and Strube 2006) to wordsenses (Fellbaum, Grabowski, and Landes 1997; Bruce and Wiebe 1998; Véronis 1998) to dialogue acts (Carletta et al. 1997; Core and Allen 1997) we find substantial differences in the distribution of tags. For these reasons the S coefficient has never really found much use in CL, and studying it does not contribute to the points we develop in this article, so we will not discuss it further.

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 said above, 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.[3]

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

$\hat{P}(k)$, 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.[4]

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

The value of π for the experiment in Table 1 is calculated as follows.

$$P(\text{Stat} \mid \text{Coder A}) = P(\text{Stat} \mid \text{Coder B}) = \hat{P}(\text{Stat}) = 0.35$$

$$P(\text{IReq} \mid \text{Coder A}) = P(\text{IReq} \mid \text{Coder B}) = \hat{P}(\text{IReq}) = 0.65$$

$$A_e^\pi = 0.35^2 + 0.65^2 = 0.1225 + 0.4225 = 0.545$$

$$\pi = \frac{0.7 - 0.545}{1 - 0.545} = \frac{0.155}{0.455} \approx 0.341$$

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.

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

3 The same method is used to compute the K coefficient discussed by Siegel and Castellan (1988), which is why we consider K to be a generalization of π rather than κ; this has already been pointed out by Di Eugenio and Glass (2004).

4 We should note that A_e^π is a **biased estimator** which overestimates the expected agreement. This is because A_e^π is calculated from a single sample, and items in a sample tend to be somewhat closer together than items in the entire population (which amounts to the loss of one "degree of freedom"). Thus, while $\hat{P}(k)$ is an unbiased estimator of the distribution of items in the entire population, A_e^π is a biased estimator of the expected agreement in the entire population; an unbiased estimator would be $(2\mathbf{i}A_e^\pi - 1)/(2\mathbf{i} - 1)$, as used for Krippendorff's α (section 2.6).

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 $\hat{P}(k|c)$, the proportion of items actually assigned by coder c_i to category k; this is the number of assignments to k by c, $\mathbf{n}_{ck}$, 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 $k \in K$.[5]

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

The value of κ for the experiment in Table 1 is as follows.

$$P(\text{Stat} \mid \text{Coder A}) = 0.3 \quad P(\text{Stat} \mid \text{Coder B}) = 0.4$$

$$P(\text{IReq} \mid \text{Coder A}) = 0.7 \quad P(\text{IReq} \mid \text{Coder B}) = 0.6$$

$$A_e^\kappa = 0.3 \times 0.4 + 0.6 \times 0.7 = 0.12 + 0.42 = 0.54$$

$$\kappa = \frac{0.7 - 0.54}{1 - 0.54} = \frac{0.16}{0.46} \approx 0.348$$

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.

What is measured by Pi and Kappa. The difference between π and κ has been the subject of much contention in the literature, both in CL where it has surfaced recently (Di Eugenio and Glass 2004; Craggs and McGee Wood 2005) and in other fields where it constitutes a long-standing debate (Byrt, Bishop, and Carlin 1993; Fleiss 1975; Krippendorff 1978, 2004b). We will discuss this difference in more detail in section 3.1, where we also prove that the values of π and κ get closer as the number of coders grows. At this point we only wish to point out that by averaging the individual distributions, π reflects our expectations for arbitrary coders, whereas κ relates specifically to the coders who performed the annotation (since it takes their individual distributions as a basis for calculating chance agreement). When generalization is desired (as in most reliabiliy studies), π is therefore more appropriate. Generalizing to arbitrary coders introduces additional variability, so it is not surprising that $\pi \leq \kappa$ for any particular set of data, with equality obtaining when the coders are indistinguishable.

A numerical comparison of S, Pi and Kappa. Zwick (1988) provides a particularly clear illustration of the effect of differences between the coders' observed distributions (**coder**

5 Since A_e^κ is calculated from two independent samples, it is an **unbiased estimator** of the expected agreement of the two specific coders on the entire population of items.

Table 4
The effect of coder marginals on coefficient values.

		CODER A				
		STAT	IREQ	IAFA	CSFA	TOTAL
	Case 1: Marginals uniform $S = 0.467, \pi = 0.467, \kappa = 0.467$					
CODER B	STAT	0.20	–	–	0.05	0.25
	IREQ	–	0.10	0.15	–	0.25
	IAFA	–	0.15	0.10	–	0.25
	CSFA	0.05	–	–	0.20	0.25
	TOTAL	0.25	0.25	0.25	0.25	1.00
	Case 2: Marginals equal but not uniform $S = 0.467, \pi = 0.444, \kappa = 0.444$					
CODER B	STAT	0.20	0.10	0.10	–	0.40
	IREQ	0.10	0.10	–	–	0.20
	IAFA	0.10	–	0.10	–	0.20
	CSFA	–	–	–	0.20	0.20
	TOTAL	0.40	0.20	0.20	0.20	1.00
	Case 3: Marginals unequal $S = 0.467, \pi = 0.460, \kappa = 0.474$					
CODER B	STAT	0.20	0.05	0.05	0.10	0.40
	IREQ	–	0.10	0.05	0.05	0.20
	IAFA	–	0.05	0.10	0.05	0.20
	CSFA	–	–	–	0.20	0.20
	TOTAL	0.20	0.20	0.20	0.40	1.00

marginals) on the values of S, κ, and π. We reproduce one of her examples here, recasting her example in a CL setting.

Let us assume a dialogue act classification scheme, again adapted from DAMSL, with four categories for forward-looking function: `statement`, `info-request`, `influencing-addressee-future-action`, and `committing-speaker-future-action`. Let us again assume we have two coders, coder A and coder B. In Table 4 we find three illustrations of the three situations that may arise, in all of which the observed agreement $A_o = 0.60$.

Case 1 is an example of the case in which the coders assign equal proportions of items to all categories; in this case, all three coefficients of agreement have the same value. Case 2 exemplifies the situation in which coder A and coder B, while not assigning equal proportions of items to all categories, still end up assigning items to

categories in identical proportions: both judge 40% of items to be `statement`, 20% to be `info-request`, and so forth. In this situation, κ and π still have the same value. Finally, Case 3 is an example of the situation in which Coder A and Coder B do not even agree on the proportion of items belonging to a given category: in this case, κ and π may have different values. Notice also that in Case 2 we get lower values of κ and π than in Case 3 – that is, when observed agreement is held constant, agreement on the marginals results in lowered coefficient values (Feinstein and Cicchetti 1990; Cicchetti and Feinstein 1990; Di Eugenio and Glass 2004).

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. The coefficients π and κ, presented above with their original definitions for two coders, can be generalized to reliability studies with more than two coders. Due to historical accident, the terminology here becomes 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.

We will call Fleiss's coefficient multi-π, reserving the name multi-κ for a proper generalization of Cohen's κ (Davies and Fleiss 1982). We will drop the multi- prefixes when no confusion is expected to arise.

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, since inevitably there will be items on which some coders agree and others disagree. We therefore measure *pairwise agreement*: Fleiss (1971) defines 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.

Another problem with multiple coders is that when the number of coders **c** is greater than two, judgments cannot be shown in a contingency table like Table 1 or Table 2, since each coder has to be represented in a separate dimension. 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 5 is an example of such an agreement table, in which the same 100 utterances from Table 1 are labeled by three coders instead of two. Di Eugenio and Glass (2004, page 97) note that compared to contingency tables like Tables 1 and 2, agreement tables like Table 5 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.)

Table 5
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)

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 5, $\mathbf{n}_{\text{Utt}_1\text{Stat}} = 2$ 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 5, we find the agreement value for Utterance 1.

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

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

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

(Notice that the above definition of observed agreement is equivalent to the mean of the two-coder obeserved 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, i.e., as the probability that any pair of judgments for an item would be in agreement – or, said otherwise, the probability that two 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 by a single distribution which reflects the combined judgments of all coders, meaning that expected agreement is calculated using $\hat{\text{P}}(k)$, 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 $\mathbf{c}$.

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

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 $k \in K$.[6]

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

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

Multi-κ. It is fairly straightforward to adapt Fleiss's proposal to generalize Cohen's κ proper to more than two coders; the development below is our own, but an identical proposal can be found in Davies and Fleiss (1982).

For multi-κ, we calculate a separate probability distribution for each annotator: the probability of assigning an item to category k by coder c is the observed proportion of such assignments $\hat{P}(k|c)$, which is the number of such assignments $\mathbf{n}_{ck}$ divided by the number of items $\mathbf{i}$.

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

The probability that two arbitrary coders assign an item to a particular category $k \in K$ is the joint probability of each coder making this assignment independently. The joint probability for two particular coders c_m and c_n is $\hat{P}(k|c_m)\hat{P}(k|c_n)$, and since all coders judge all items, the joint probability for an arbitrary pair of coders is the arithmetic mean of $\hat{P}(k|c_m)\hat{P}(k|c_n)$ over all coder pairs c_m, c_n. Again, the expected agreement is the sum of this joint probability over all the categories $k \in K$.

$$A_e^\kappa = \sum_{k \in K} \frac{1}{\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \hat{P}(k|c_m)\hat{P}(k|c_n)$$

It is easy to see that A_e^κ for multiple coders is the mean of the two-coder A_e^κ values from section 2.4 for all coder pairs.

2.6 Krippendorff's α and other weighted agreement coefficients

A serious limitation of both π and κ is that they treat all disagreements equally. For some coding tasks, however, disagreements are not all alike. Even in the simpler 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 yield extremely low values and are thus not very useful; instead, what is needed are coefficients that can take into account the magnitude of the disagreements.

6 As in the two-coder case, multiple-coder A_e^π is a biased estimator calculated from a single sample; an unbiased estimator would be $(\mathbf{ic}A_e^\pi - 1)/(\mathbf{ic} - 1)$.

In this section we discuss α (Krippendorff 1980, 2004a) – a coefficient defined in a general way that is appropriate for use with multiple coders, different magnitudes of disagreement, and also missing values, and 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. It is 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 idenical 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 1970b).

Variance is a useful concept if the coders assign numerical values to the items (as in magnitude estimation tasks). We follow the standard definition of a sample's variance s^2 as the sum of square differences from the mean $SS = \sum(x - \bar{x})^2$ divided by the degrees of freedom df. Each item in a reliability study can be considered to be a separate level in a single-factor analysis of variance: the smaller the variance around each level, the higher the reliability. In order to be comparable across studies, the variance within the levels (s^2_{within}) needs to be scaled with respect to the expected variance, which is estimated by the overall variance of the data (s^2_{total}). The ratio s^2_{within}/s^2_{total} has the following properties.

- $s^2_{within}/s^2_{total} = 0$ when agreement is perfect (no variance within the levels).
- $s^2_{within}/s^2_{total} = 1$ when agreement is the result of chance.
- $s^2_{within}/s^2_{total} > 1$ when there is systematic disagreement.

Subtracting the ratio s^2_{within}/s^2_{total} from 1 gives a coefficient with the same "anchors" as the ones from the previous sections, namely the value 1 signifies perfect agreement while 0 signifies chance agreement.

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

We also note that the ratio s^2_{within}/s^2_{total} cannot exceed 2: $SS_{within} \leq SS_{total}$ by definition, and $df_{total} < 2\,df_{within}$ because each item has at least two judgments. The lower bound for α is therefore -1.

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 since 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 judgments 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 reference 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 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; note also that the formula below incorrectly counts the number of pairs of identical judgments, but there's no need to correct for this 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 do 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, D_e^α, is twice the total variance.

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

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

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

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 the fact that π uses a biased estimator of the expeceted 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 ($\mathbf{d}_{ab} = 0$ 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 $\mathbf{d}_\text{max}$.

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

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: $\text{D}_\text{o}^{\kappa_w} = 1 - \text{A}_\text{o}^{\kappa}$.

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 $\hat{\text{P}}(k|c)$, 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 $\mathbf{i}$.

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

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 $\hat{\text{P}}(k_a|c_1)\hat{\text{P}}(k_b|c_2)$. 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.

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

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: $\text{D}_\text{e}^{\kappa_w} = 1 - \text{A}_\text{e}^{\kappa}$.

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.

$$\kappa_w = 1 - \frac{\text{D}_\text{o}}{\text{D}_\text{e}}$$

2.7 The Coefficient Cube

The agreement coefficients we have seen can all be thought of as modifications of Scott's π along three different dimensions. One dimension is the calculation of expected

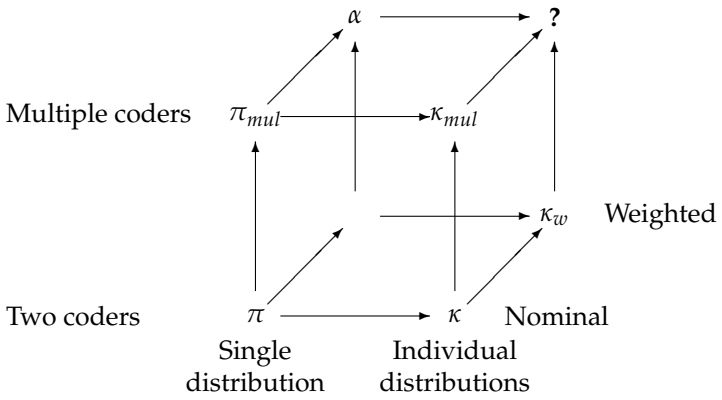

Figure 1
Generalizing π along three dimensions

agreement using separate probability distributions for the individual coders, as done by κ. Another dimension is a generalization of the original two-coder definitions to multiple coders, resulting in multi-π (Fleiss's κ) and multi-κ (Davies and Fleiss 1982). A third dimension is the introduction of weighted agreement coefficients – α for multiple coders with a single distribution, and κ_w for two coders with separate distributions. The relations between the various coefficients are depicted in Figure 1.

What is missing from the picture is a coefficient that generalizes π along all three dimensions – an agreement coefficient that is weighted, applies to multiple coders, and calculates expected agreement using a separate probability distribution for each coder. Such a coefficient can be thought of as a generalization of κ_w to multiple coders, or alternatively as a modification of α which uses individual coders' distributions for determining chance agreement. We now develop such a coefficient, calling it α_κ which should serve as a reminder that it is shares properties of both κ and α. (In a previous version of this paper we called this coefficient β).

Like the other weighted coeficients, α_κ measures the observed and expected disagreement, whose ratio is subtracted from one.

$$\alpha_\kappa = 1 - \frac{D_o}{D_e}$$

The observed disagreement is the same as for the other weighted measures, that is the mean disagreement per item, where the disagreement per item is the mean distance between all the judgment pairs pertaining to it (section 2.6).

The expected disagreement is the expected distance for an arbitrary judgment pair, which is the arithmetic mean of all possible distances between category pairs weighted by the probabilities for choosing particular pairs. We estimate the probability that coder c assigns an item to category k as the total number of such assignments $\mathbf{n}_{ck}$ divided by the overall number of assignments for this coder, which is the number of items $\mathbf{i}$.

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

The probability that two particular coders c_m and c_n assign an item to two distinct categories k_a and k_b is $\hat{P}(k_a|c_m)\hat{P}(k_b|c_n) + \hat{P}(k_b|c_m)\hat{P}(k_a|c_n)$. Since all coders judge all

items, the probability that an arbitrary pair of coders assign an item to k_a and k_b is the arithmetic mean of $\hat{P}(k_a|c_m)\hat{P}(k_b|c_n) + \hat{P}(k_b|c_m)\hat{P}(k_a|c_n)$ over all coder pairs.

$$\begin{aligned}\hat{P}(k_a, k_b) &= \frac{1}{\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \hat{P}(k_a|c_m)\hat{P}(k_b|c_n) + \hat{P}(k_b|c_m)\hat{P}(k_a|c_n) \\ &= \frac{1}{\mathbf{i}^2\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \mathbf{n}_{c_m k_a}\mathbf{n}_{c_n k_b} + \mathbf{n}_{c_m k_b}\mathbf{n}_{c_n k_a}\end{aligned}$$

The expected disagreement is the mean of the distances for all distinct category pairs, weighted by the above probabilities (recall that identical category pairs contribute a distance of zero, so it does not matter if and how they are counted).

$$\begin{aligned}D_e^{\alpha_\kappa} &= \sum_{j=1}^{\mathbf{k}-1} \sum_{l=j+1}^{\mathbf{k}} \hat{P}(k_j, k_l)\mathbf{d}_{k_j k_l} \\ &= \frac{1}{\binom{\mathbf{c}}{2}} \sum_{j=1}^{\mathbf{k}-1} \sum_{l=j+1}^{\mathbf{k}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \Big(\hat{P}(k_j|c_m)\hat{P}(k_l|c_n) + \hat{P}(k_l|c_m)\hat{P}(k_j|c_n)\Big)\mathbf{d}_{k_j k_l} \\ &= \frac{1}{\binom{\mathbf{c}}{2}} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \hat{P}(k_j|c_m)\hat{P}(k_l|c_n)\mathbf{d}_{k_j k_l} \\ &= \frac{1}{\mathbf{i}^2\binom{\mathbf{c}}{2}} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \mathbf{n}_{c_m k_j}\mathbf{n}_{c_n k_l}\mathbf{d}_{k_j k_l}\end{aligned}$$

It is easy to see that $D_e^{\alpha_\kappa}$ is the mean of the $D_e^{\kappa_w}$ values (section 2.6) over all coder pairs. If we take all disagreements to be of equal weight, that is $\mathbf{d}_{k_a k_b} = 1$ for all $k_a \neq k_b$, then this measure of expected disagreement is exactly the complement of the expected agreement for multi-κ as calculated in section 2.5: $D_e^{\alpha_\kappa} = 1 - A_e^{\kappa}$.

2.8 An Integrated Example

We end this section with an example illustrating how all of the agreement coefficients discussed above are computed. To facilitate comparisons, all computations will be based on the annotation statistics in Table 6. This confusion matrix reports the results of an experiment where two coders classify a set of utterances into three categories.

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 figures on the diagonal of the confusion matrix in Table 6) and dividing by the total number of items.

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

Table 6
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

Expected agreement for S is the reciprocal of the number of categories, or $\frac{1}{3}$; S is the observed agreement, discounted by this fraction.

$$A_e^S = \frac{1}{3}$$

$$S = \frac{A_o - A_e^S}{1 - A_e^S} = \frac{0.88 - \frac{1}{3}}{1 - \frac{1}{3}} = 0.82$$

Expected agreement for π is the sum over all categories of the square of the mean of the individual coders' proportions; π is the observed agreement, discounted by this value.

$$A_e^\pi = \left(\frac{46+52}{2\times 100}\right)^2 + \left(\frac{44+32}{2\times 100}\right)^2 + \left(\frac{10+16}{2\times 100}\right)^2 = 0.49^2 + 0.38^2 + 0.13^2 = 0.4014$$

$$\pi = \frac{A_o - A_e^\pi}{1 - A_e^\pi} = \frac{0.88 - 0.4014}{1 - 0.4014} \approx 0.7995$$

Expected agreement for κ is the sum over all categories of the product of the individual coders' proportions; κ is the observed agreement, discounted by this value.

$$A_e^\kappa = \frac{46}{100}\times\frac{52}{100} + \frac{44}{100}\times\frac{32}{100} + \frac{10}{100}\times\frac{16}{100} = 0.396$$

$$\kappa = \frac{A_o - A_e^\kappa}{1 - A_e^\kappa} = \frac{0.88 - 0.396}{1 - 0.396} \approx 0.8013$$

We see that the values of π and κ are very similar, which is to be expected when agreement is high, since 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.

Weighted coefficients. Suppose we notice that while `Statements` and `Info-Requests` are clearly distinct classifications, `Checks` are 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–Statement:	0	Statement–Info-Request:	1
Info-Request–Info-Request:	0	Statement–Check:	0.5
Check–Check:	0	Info-Request–Check:	0.5

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 calculation below 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 6 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.

Expected disagreement for α is the sum over all category pairs of the product of the sum of the individual coders' judgments, weighted by the distance and by the total number of items 2×100 times the degrees of freedom $2 \times 100 - 1$; α is the observed disagreement, discounted by this value and subtracted from 1.

$$\begin{aligned} D_e^{\alpha} = {} & \tfrac{(46+52)\times(46+52)}{2\times100\times(2\times100-1)} \times 0 + \tfrac{(44+32)\times(46+52)}{2\times100\times(2\times100-1)} \times 1 + \tfrac{(10+16)\times(46+52)}{2\times100\times(2\times100-1)} \times 0.5 \\ & + \tfrac{(46+52)\times(44+32)}{2\times100\times(2\times100-1)} \times 1 + \tfrac{(44+32)\times(44+32)}{2\times100\times(2\times100-1)} \times 0 + \tfrac{(10+16)\times(44+32)}{2\times100\times(2\times100-1)} \times 0.5 \\ & + \tfrac{(46+52)\times(10+16)}{2\times100\times(2\times100-1)} \times 0.5 + \tfrac{(44+32)\times(10+16)}{2\times100\times(2\times100-1)} \times 0.5 + \tfrac{(10+16)\times(10+16)}{2\times100\times(2\times100-1)} \times 0 \end{aligned}$$

$$= \frac{1}{39800} \times (2 \times 98 \times 76 + 2 \times 98 \times 26 \times 0.5 + 2 \times 76 \times 26 \times 0.5) \approx 0.4879$$

$$\alpha = 1 - \frac{D_o}{D_e^{\alpha}} \approx 1 - \frac{0.09}{0.4879} \approx 0.8156$$

Finally, expected disagreement for α_κ is the sum over all category pairs of the products of the individual coders' proportions, weighted by the distance; α_κ is the observed disagreement, discounted by this value and subtracted from 1.

$$\begin{aligned} D_e^{\alpha_\kappa} = {} & \tfrac{46}{100} \times \tfrac{52}{100} \times 0 + \tfrac{44}{100} \times \tfrac{52}{100} \times 1 + \tfrac{10}{100} \times \tfrac{52}{100} \times 0.5 \\ & + \tfrac{46}{100} \times \tfrac{32}{100} \times 1 + \tfrac{44}{100} \times \tfrac{32}{100} \times 0 + \tfrac{10}{100} \times \tfrac{32}{100} \times 0.5 \\ & + \tfrac{46}{100} \times \tfrac{16}{100} \times 0.5 + \tfrac{44}{100} \times \tfrac{16}{100} \times 0.5 + \tfrac{10}{100} \times \tfrac{16}{100} \times 0 \end{aligned}$$

$$= 0.49$$

$$\alpha_\kappa = 1 - \frac{D_o}{D_e^{\alpha_\kappa}} = 1 - \frac{0.09}{0.49} \approx 0.8163$$

3. Bias and prevalence

Two issues with agreement coefficients, recently brought to light by Di Eugenio and Glass (2004), concern the behavior of the 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 to understand the differences between the coefficients.

3.1 Annotator bias

The difference between the coefficients on the 'left' face and the 'right' face of the coefficient cube (Figure 1 in section 2.7) 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 (Byrt, Bishop, and Carlin 1993; Craggs and McGee Wood 2005; Di Eugenio and Glass 2004; Fleiss 1975; Hsu and Field 2003; Krippendorff 1978, 2004b; Zwick 1988).

A claim often repeated in the literature is that single-distribution coefficients like π and α are based on the assumption 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. Thus, 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 response to this, 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. However, 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. Both π-style and κ-style coefficients assume that the annotators code the data according to properties inherent in the data, and that variation arises from various sources, some systematic and some arbitrary. The difference is only in the understanding of the notion of "chance agreement". Therefore, regardless of how divergent the actual coders are, both kinds of coefficients are applicable; they just differ in meaning.

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). We feel, however, that the bias problem is less paradoxical than it sounds. While 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 observed agreement and expected agreement 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 the expected agreement. 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 while 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, and 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 6 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; a formal proof is given by Artstein and Poesio (2005) and repeated below.[7]

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. Note that annotator bias is not related to sampling bias (the source of difference between π and α), nor is it the same as the Bias Index BI of Byrt, Bishop, and Carlin (1993). 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 $\hat{P}(k|c)$ for all categories $k \in K$, divided by the number of coders $\mathbf{c}$ less one (shown as part of the proof in appendix A).

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

7 Craggs and McGee Wood (2005) also suggest increasing the number of coders in order to overcome individual annotator bias, but without proof.

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

$$\kappa - \pi = \frac{A_o - (A_e^\pi - B)}{1 - (A_e^\pi - B)} - \frac{A_o - A_e^\pi}{1 - A_e^\pi} = B \cdot \frac{(1 - A_o)}{(1 - A_e^\kappa)(1 - A_e^\pi)}$$

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

1. For any particular coding data, $A_e^\pi \geq A_e^\kappa$, because B is the sum of non-negative numbers.
2. For any particular coding data, $\kappa \geq \pi$, because the difference between them is the product of non-negative numbers.
3. The difference between κ and π grows as the annotator bias grows: for a constant A_o and A_e^π, a greater B implies a greater value for $\kappa - \pi$.

It is also easy to show that the following holds:

Observation. *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* $\hat{P}(k|c)$ *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. To show this we define a coefficient α_b, which is just like α except that it uses a biased estimator for expected disagreement, that is $D_e^{\alpha_b} = (\mathbf{ic} - 1)D_e^\alpha / \mathbf{ic}$. Now, for any particular coding data, $D_e^{\alpha_\kappa} \geq D_e^{\alpha_b}$, and consequently $\alpha_\kappa \geq \alpha_b$; the greater the number of coders, the lower the difference between α_κ and α_b (for proof see appendix B). It is easy to see that α and α_b approach each other as either the number of items or the number of coders grows, and therefore α and α_κ also converge with more coders. This means that the more coders we have, the less important the choice of coefficient among α, α_b, and α_κ. In an annotation study with 18 subjects (Poesio and Artstein 2005) we calculated all three coefficients and found that their values never differed beyond the third decimal point: for example, we found $\alpha = 0.69115$, $\alpha_b = 0.69091$, and $\alpha_\kappa = 0.69111$ for the condition of full chains with Dice distance metric (see section 4.4 for an explanation of the various conditions).

In summary, our views concerning the difference between π-style and κ-style coefficients are as follows: (i) Reporting two coefficients, as suggested by Di Eugenio and Glass (2004), is unlikely to help. Instead, 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, since Siegel and Castellan's K is identical to π. If the coefficient is used in order to assess the correctness of data points agreed upon by two coders (or more), then Cohen's κ or its generalizations κ_w or α_κ may be more appropriate. (ii) However, the numerical difference between single-distribution coefficients (π) and individual-distribution coefficients (κ) is often not very large, especially in cases of high agreement. (iii) Further, the numerical difference decreases as the number of annotators grows.

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

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 observed agreement is very high, yet chance-corrcted agreement is low (Feinstein and Cicchetti 1990; Cicchetti and Feinstein 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.

This surprising result is, however, correct and 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 the observed agreement $A_o = 1-2\epsilon$ and the expected agreement $A_e = 1-2(\delta+\epsilon)+2(\delta+\epsilon)^2$, 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 approximatly $\delta/(\delta+\epsilon)$, that is the value we get if we take all the items marked by one particular coder as **Rare**, and calculate what proportion of those items were labeled

Rare by the other coder. This is indeed a measure of the ability to agree on the rare category, so it is a good measure of reliability.

We therefore do not agree with the recommendation of Di Eugenio and Glass (2004) to report an additional coefficient when one of the categories is very common (Di Eugenio and Glass recommend reporting $2A_o - 1$, which is the value of S when there are exactly two categories). If reliability is a concern, then the appropriate reliability coefficient should be reported, and S does not reflect reliability precisely because it is insensitive to the difference between common and rare categories. If reliability turns out to be low but it is still of interest to note that overall agreement was high, then it is best to report the raw observed agreement A_o, since this value is easier to interpret than S. The reporting of raw agreement figures should be accompanied by a note explaining that these figures are not corrected for chance and therefore do not reflect reliability.

4. Using agreement measures for CL annotation tasks

We will now review the use of intercoder agreement measures in CL ever since Carletta's original paper in the 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, focusing in particular on the types of annotation where there has been some debate concerning the most appropriate measures of agreement. To our knowledge the relative merits of biased versus unbiased measures have only been discussed by Di Eugenio and Glass (2004) and Craggs and McGee Wood (2005), but there has been some debate concerning the use and merit of weighted coefficients.

We will also try to highlight examples of good practice. Krippendorff (2004a, chapter 11) bemoans 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 in CL, such as the creation of the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 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). We are not aware of any annotation effort in CL that applies a methodology as rigorous as that envisaged by Krippendorff and discusses it next to a study of the reliability of their coding scheme, but we will highlight a few studies that are particularly sound, focusing on the methodology and the coefficients used rather than on the scores.

4.1 Methodological recommendations from Content Analysis

An extensive discussion of the methodology to be followed in carrying out a reliability study can be found in chapter 11 of Krippendorff (2004a). We summarize here his main recommendations as a preliminary for the discussion of CL practice.

4.1.1 Generating data to measure reproducibility. Krippendorff's recommendations are intended to apply to the field of Content Analysis, where coding is a preliminary step 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 as 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 the above requirements. The first requirement is violated by the practice of expanding the written coding instructions and including new rules as the data get 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 reproducibility data 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.

Table 8
Kappa values and strength of agreement according to Landis and Koch (1977).

KAPPA VALUES	STRENGTH OF AGREEMENT
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Perfect

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 Table 8, 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 enormously influential paper, CL researchers have attempted to achieve a value of K (more seldom, of α) above the magical 0.8 threshold, or, failing that, the 0.67 level allowing for "tentative conclusions". However, we should point out that 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), analyzing several proposals concerning 'acceptable' reliability, concludes that "reliability coefficients of .9 or greater would be acceptable to all, .8 or greater would be acceptable in most situations, and below that, there exists great disagreement."

This is clearly a fundamental issue – there is little point in running a reliability study if then we can't interpret the results, that is, decide whether we have reached enough agreement for our purposes – but as we will see in the rest of this section, practical experience with using these coefficients in CL hasn't helped in settling the matter. In fact, we will see in the rest of this section that 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 more complicated. We will return to the issue of interpreting the value of the coefficients at the end of this article.

4.2 Labeling units with a common and predefined set of categories

The most basic 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 equidistant; this assumption seems to have worked out well in practice, but we will also consider studies that question it.

4.2.1 Part-of-speech tagging. The simplest type of linguistic annotation is the annotation of parts of speech. Historically, this was also the first type of annotation to be carried out over a 1-million-word corpus, the Brown corpus (Francis and Kucera 1982), and then over a 100-million-word corpus, the British National Corpus (Leech, Garside, and Bryant 1994).[8] None of these early efforts involved systematic tests of the reliability of the annotation. Such studies were however carried out for later efforts, such as the annotation of the TIGER corpus of German (Brants and Plaehn 2000), in which however only percentage agreement was computed. Agreement studies using chance-corrected measures were carried out for the annotation of the GENIA corpus (Tateisi and Tsuji 2004) and by Mieskes and Strube (2006), among others. In both these studies an unweighted, unbiased measure was used, K. These studies generally report very high levels of agreement, particularly for so-called 'interactive' mode of annotation where annotators correct the output of an automatic POS tagger, pioneered by the BNC

8 http://www.natcorp.ox.ac.uk/docs/gramtag.html

annotation and which is at the moment the standard method for this type of annotation (for example, Mieskes and Strube report K = 0.96 for this mode).

This considerable experience with POS annotation gives the field confidence that current tagsets are adequate for the purpose of creating large-scale POS-annotated corpora, at least for English. It is worth noting however that it not clear from the literature whether any of these agreement studies satisfies the three requirements laid out by Krippendorff. It is equally clear that from a linguistic perspective, treating all distinctions between POS tags as having the same weight is a considerable simplification. For instance, even a coarse-grained tagset like the Penn Treebank POS tagset makes a distinction between singular and plural nouns; yet intuitively, disagreeing on whether, say, a particular instance of the word *deer* should be tagged as a plural noun (tag `NNS`) or a singular one (tag `NN`) is less of a disagreement than disagreeing on whether, say, a particular word should be classified as a determiner or a noun. This intuition is supported by the analysis carried out in the one detailed study of human performance at POS tagging we are aware of, by Babarczy, Carroll, and Sampson (2006). In analyzing the disagreements between two (highly experienced) annotators using the SUSANNE part of speech tagset, the authors examined three types of disagreement: **fine** disagreement, **coarse** disagreement, and **major parts of speech** disagreement. Fine disagreement is the case in which the annotators chose two distinct POS tags from the Susanne tagset (which contains around 300 tags). Coarse disagreements are those obtained when using the simplified tagset that has been used for automatic annotation, obtained by collapsing some of the tags by removing the final character of the tag label (for example, replacing `NN1c` – the SUSANNE tag for singular count nouns – with `NN1`; this left around 180 categories). Finally, major parts of speech disagreements are those between categories such as `N` (noun), `V` (verb), etc. (only 18 labels). Babarczy, Carroll, and Sampson found, naturally, that percentage agreement was greater over major parts of speech (98.5%), slightly lower (98.0%) in the case of coarse comparisons, and lower still in the case of fine comparisons (97.4%). No chance-corrected agreement results were reported. Moreover, Babarczy, Carroll, and Sampson observed that many of the disagreements were among close categories: for example, 32% of the disagreements involved classification of proper names (the fine-grained version of the SUSANNE scheme requires making distinctions between surnames and organizations, for instance), and 15.8% of the disagreements were caused by the coders disagreeing on whether words like *training* should be tagged as nouns or participles in contexts like noun-noun compounds (e.g., *training centre*). As we will see, this idea of identifying and separating levels of annotation requiring progressively more complex decisions has been proposed for many types of annotation tasks.

4.2.2 Dialogue act tagging. Dialogue act tagging is another type of linguistic annotation with which by now the CL community has had extensive experience. Dialogue-act-annotated spoken language corpora include MapTask (Carletta et al. 1997), Switchboard (Stolcke et al. 1997), 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 originated 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. 1997), 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 underlied the tests of agreement on part of speech tagging discussed earlier: that every unit (utterance) is assigned to exactly one category (dialogue act), and that these categories are distinct (the number of dialogue act tags tends to be much smaller than the number of POS tags). Therefore, again, unweighted measures, and in particular K, tend to be used to measure inter-coder agreement.[9]

However, a rather more serious challenge to these assumptions has arisen in the case of dialogue act tagging, from theories of dialogue acts based on the observation that utterances tend to have more than one function at the dialogue act level (Traum and Hinkelman 1992; Bunt 2000; Allen and Core 1997); 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 some systems, the fact that an utterance is performed to 'answer' a particular question is also marked – for example, by tagging it as a `reply-XX` in the MapTask coding scheme, as expressing a 'backward communicative function' in DAMSL (Allen and Core 1997), or by using an 'answerhood' rhetorical relation in the system of Traum and Hinkelman. 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 (see for example the `explain` and `acknowledge` tags in the MapTask coding scheme). By contrast, in the annotation schemes inspired from these newer theories such as DAMSL (Allen and Core 1997), coders are allowed to assign tags at different levels.

This solution also addresses another problem with the older schemes: for in addition to performing dialogue acts at different levels, utterances can also perform multiple functions at the very same level – for example, the core level (in DAMSL parlance, they can perform more than one 'forward communication function'). Utterances such as (1), for instance, perform both what in DAMSL would be called a `info-request` dialogue act and some sort of suggestion (called `open-option` in DAMSL), or perhaps even a `directive`.

(1) Could we meet at the Wigmore Hall at 11 am?

So-called 'checks' might be viewed as another example of utterances performing multiple core functions: For example, utterances 5.4–5.6 below might be viewed as expressing both a `statement` and an `info-request`: M is at the same time stating his or her belief that one boxcar of oranges is sufficient to make a tanker, and requesting S to confirm this belief (TRAINS 1991 [Gross, Allen, and Traum 1993], dialogue d91-2.2).[10]

(2)
```
5.4  M: I assume one
5.5     one boxcar
5.6     of oranges is enough to make a tanker
6.1  S: yeah
```

9 To our knowledge, Carletta et al. (1997) were the only group among those carrying out these early studies who considered using α for measuring agreement on dialogue act annotation.

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

In the MapTask scheme, checks are treated as a separate class of dialogue acts, which may be one of the reasons why the number one source of confusion found by Carletta et al. was that between the tags `check` and `query-yn`; collapsing all question-type tags into one resulted in an increase of the agreement from $K = 0.83$ to $K = 0.89$.[11] In DAMSL, by contrast, coders are allowed to annotate more than one forward communicative function.

Two annotation experiments 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 (Forward and Backward) 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 (1997) 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. (1998) observed higher agreement, ranging from $K = 0.93$ (for `other-forward-function`) to 0.54 (for the backward function `agreement`).

Core and Allen found that many disagreements resulted from some of the coders choosing different subsets of communicative functions: for example, one of the main sources of disagreement was the difficulty for coders to tell whether an utterance is an acceptance or simply an acknowledgment, two types of Backward Communicative Function treated as separate dimensions. They also observed a problem with checks: some coders would mark an utterance like 5.4–5.6 in (2) as having both a `statement` and an `info-request` function, whereas others would mark it as expressing only a `statement`, or only an `info-request`. Because agreement was measured for each communicative function independently, partial agreement on the overall 'label' (the entire set of labels assigned to an utterance in all dimensions) could not be taken into account in such cases. In cases in which annotations overlapped (as in the case of an utterance which one of the coders marked both a `statement` and an `info-request`, whereas the other coder only marked an `info-request`), Core and Allen would get perfect agreement along one dimension (`statement`), but no agreement at all along the `Influencing-Addressee-Future-Action` dimension. It might be argued that in such cases, using a weighted coefficient measuring agreement over the entire set of labels assigned to an utterance in all dimensions might provide a better indication of the actual agreement on the interpretation of that utterance.

These problems led many researchers to return to 'flat' tagsets for dialogue acts after experimenting with multidimensional ones, 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 SWITCHBOARD-DAMSL by Jurafsky, Shriberg, and Biasca (1997), who annotated the Switchboard corpus in order to study the interaction of dialogue acts and speech recognition (Stolcke et al. 1997). Jurafsky, Shriberg, and Biasca started by running an annotation pilot using the DAMSL scheme. They found that only 220 of the possible combination of tags occurred in the corpus, but also that agreement was not very high. A new tagset called SWITCHBOARD-DAMSL was then developed consisting of only 42 tags, on which good agreement was

11 Specifically, checks are defined as "[Questions asking] for confirmation of material which the speaker believes might be inferred, given the dialogue context" – Carletta et al. (1997, Figure 1). Carletta et al. point out that in practice the coders used the `check` tag to mark utterances querying information that the speaker believed had been told, as in: G: *Ehm, curve round slightly to your right.* F: *to my right?*

found. This new tagset 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. Similarly, Doran et al. (2001) decided not to adopt the DAMSL scheme on grounds of complexity (without running a pilot), adopting instead a simplified version of the tagset developed by the CSTAR consortium, in part because it seemed more appropriate for the task.

Interestingly, subsequent developments of SWITCHBOARD-DAMSL backtracked on some of these decisions. 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. only achieved a comparable reliability to that obtained with SWITCHBOARD-DAMSL when using a tagset of only five 'class-maps'. This aspect of the ICSI-MRDA was further developed in the MALTUS scheme proposed by Popescu-Belis (2005), in which further constraints are introduced in the composition of class maps so as to greatly reduce the number of theoretically possible multi-labels from around 7 million to around 200.

In addition, Shriberg et al. (2004) also introduced a hierarchical organization of tags to improve reliability. The dimensions of the DAMSL scheme can be viewed as 'super-classes' 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` mentioned earlier, 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 such as (1) 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 measures of agreement – specifically, κ_w – for measuring agreement over their hierarchical tagging scheme. There are a number of problems with the Geertzen and Bunt proposal, ranging from the hierarchy they propose to the equations given in the paper for computing the distance metric $\mathbf{d}$, but we feel nevertheless that the work is worth discussing as one of the few examples of use of weighted measures of agreement in CL.

Geertzen and Bunt were testing 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 to allow coders to leave some aspects of the coding undecided. For example, the difficult case repeatedly mentioned in this section, `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`. This organization is shown in Figure 2.

The distance metric $\mathbf{d}$ 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.

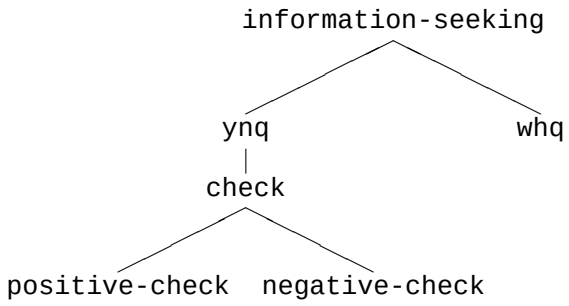

Figure 2
Hierarchical tags from Geertzen and Bunt (2006)

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 result is the following distance metric.[12]

$$\mathbf{d}(c_i, c_j) = 1 - h(c_i, c_j) \times a^{\Delta(c_i, c_j)} \times b^{\Gamma(c_i, c_j)}$$

Here, $h(c_i, c_j)$ is 1 if c_i and c_j are identical or stand in an ancestor relation in the hierarchy, and 0 if they don't; $0 < a < 1$ is a constant expressing the amount of disagreement associated with a certain distance between levels in the hierarchy and $\Delta(c_i, c_j)$ is the difference in depth between the levels of c_i and c_j; and $0 < b \leq 1$ is the depth-dependent correction factor and $\Gamma(c_i, c_j)$ is the minimal depth of c_i and c_j.[13] For example, assuming the tree in Figure 2, and given the values $a = 0.75$ and $b = 1$ as in Geertzen and Bunt (2006), we get the following values for $\mathbf{d}$.

$$\mathbf{d}(\texttt{ynq}, \texttt{whq}) = 1 - 0 \times 0.75^0 \times 1 = 1$$
$$\mathbf{d}(\texttt{ynq}, \texttt{ynq}) = 1 - 1 \times 0.75^0 \times 1 = 0$$
$$\mathbf{d}(\texttt{ynq}, \texttt{check}) = 1 - 1 \times 0.75^1 \times 1 = 0.75$$

The results of an agreement test with two annotators run by Geertzen and Bunt are shown in Table 9. The first two columns give the values for κ and κ_w; the third column is the number of pairs on which the coefficients were computed; the fourth column is the AP ratio for each dimension – the proportion of cases which were marked on that dimension by both annotators out of the number of cases which were marked by at least one annotator. As the table shows, 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

12 Geertzen and Bunt define a measure of *closeness*, $\delta(c_1, c_2)$, and then modify the equation for κ_w. We changed things slightly to relate the proposal more closely to the equations seen in section 2.6.

13 $h(c_i, c_j)$ was omitted by mistake from the version of this equation in the paper in the SIGDIAL proceedings (Harry Bunt, personal communication). Also, as far as we can see, identical categories have a distance of 0 only when $b = 1$.

Table 9
A comparison of κ and κ_w agreement values for the DIT++ annotation (excerpt from Geertzen and Bunt 2006, Table 2)

Dimension	κ	κ_w	#pairs	AP ratio
contact management	1.00	1.00	8	0.17
own comm. management	1.00	1.00	2	0.08
social obl. management	1.00	1.00	61	0.80
turn management	0.82	0.82	115	0.18
dialog str. management	0.74	0.74	15	0.31
time management	0.58	0.58	68	0.72
allo feedback	0.42	0.58	17	0.14
auto feedback	0.21	0.57	127	0.34

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 elsewhere on the difficulty of interpreting the values of weighted coefficients of agreement – 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 and the best way of measuring agreement on this type of annotation was quite long, but it introduced issues that we will see discussed in the case of other CL annotation tasks as well. There are by now 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, the most prudent recommendation at this point would be to use it. The readers should however be aware that the mutual exclusivity assumption is very dubious, and that multi-dimensional or hierarchical dialogue act tagsets need not automatically result in lower reliability or in an explosion in the number of labels. If a hierarchical tagset is used, readers should be aware that weighted coefficients do capture partial agreement. But none of these decisions would be unproblematic. A hierarchical scheme designed on the basis of our intuitions about intentions 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. And once we start using weighted coefficients, interpreting the value we obtain becomes even more of a black art. We will return to both of these problems in what follows.

4.2.3 Named Entities. Named entity recognition is the task of identifying mentions of individuals and assigning them a type – e.g., finding and labeling all mentions of people or proteins in a text. It is only recently that this aspect of semantic interpretation has been identified as a separate and useful task in CL, but it has quickly grown in importance thanks to its inclusion among the information extraction subtasks first in the MUC and then in the ACE initiatives. We discuss it here as yet another example of a task originally defined as a basic labeling.

In the MUC guidelines for named entity tagging (Chinchor 1997), named entity recognition was viewed as a basic labeling task in the sense used in this section. Five

types of named entities were identified as being particularly relevant, and annotated: person, location, organization, temporal expression, and numerical expressions (e.g., "15 dollars"). We are not aware of a reliability study being carried out for the MUC annotation. With ACE (Doddington et al. 2000), the definition of the task was extended by allowing for seven types of entities instead of five (person, organization, geo-political, location, facility, vehicle, and weapon) and by introducing subtypes (such as 'building' or 'bridge' for the type 'facility').

This view of general-purpose named entity tagging as a simplified form of ontological labeling is quite natural, and efforts following ACE have tended to pursue this direction. Sekine, Sudo, and Nobata (2002) developed an Extended Named Entity Hierarchy, a taxonomy of types currently consisting of around 200 types. In named entity annotation for the biomedical domain, arguably the most common version of named entity tagging at the moment, named entity annotation according to an ontology was already the method adopted for the annotation of the GENIA corpus (Tateisi et al. 2000). Ontology-based annotation is becoming more common in connection with work on the Semantic Web (Cimiano and Handschuh 2003; Handschuh 2005) but we are not aware of any studies of agreement.

We are not aware of any study of intercoder agreement for named entity tagging which reports chance-corrected measures. The ACE 2003 annotation of the English named entity labeling task reported 88% percent intercoder agreement, whereas in the case of biomedical annotation, Tateisi et al. (2000) reported an F-measure[14] of 75.85 on a set of 4 tags (protein, dna, rna, source), and Vlachos et al. (2006) report 91% percent agreement on gene names. What makes named entity tagging interesting from the perspective of this article is that it is a simplified version of the much more complex problem of wordsense tagging discussed later. On the one hand, the categories are usually clearly distinct, which suggests that the disjointness assumption behind unweighted measures such as K may be appropriate, except perhaps in the case of metonymy (for example in *Vietnam was the source of much soul-searching in the USA*, where 'Vietnam' could refer to the country or the war). On the other hand, even more than in the case of wordsense tagging, it is natural to view the set of labels as having a taxonomic organization (e.g., depending on the annotator's knowledge, an `organization` may be classified as a `company` or as a `governmental organization` or as a `non-profit`), hence different levels of precision may be reached, which suggests that weighed coefficients of agreement may be more appropriate.

4.2.4 'Non-linguistic' coding tasks. Modern CL research is more and more concerned with extracting information that is less clearly 'linguistic' (in the traditional sense), thus overlapping more and more with the concerns of content analysis. One example of this trend is the work by Craggs and McGee Wood (2004) on annotating emotions and the work by Bruce, Wiebe and collaborators on detecting subjective judgments.

The work by Craggs and McGee Wood (2004) on developing an annotation scheme for emotions and testing its reliability has several aspects worth mentioning. First of all, there is the problem of the units to which to apply the labels. As Craggs and McGee Wood point out, 'emotional episodes' are not associated with any specific linguistic event, but persist for a certain amount of time, fading after a while. We'll return later

14 F is normally used to measure performance against a gold standard. However, it has also been used in MUC as a way of measuring agreement between two sets of results neither of which was the gold standard.

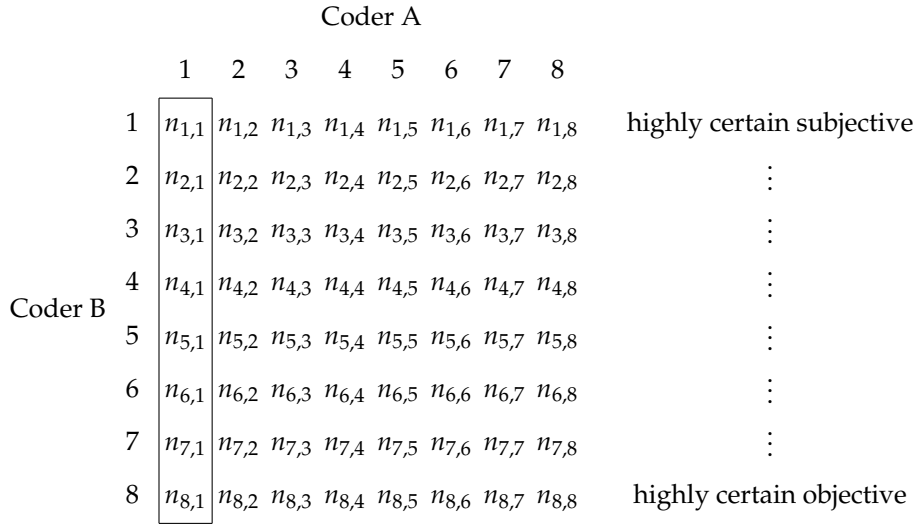

Figure 3
The profile of "highly certain subjective" for coder A is the vector of corresponding judgments by coder B.

to the issue of unitizing, but we are not aware of any clear solution to this problem. The second interesting point about this work is that it is one of the few pieces of work in which Krippendorff's α is used with a weighted distance metric, in this case, to measure distance between emotions in 'Activation-Evaluation space'. Thirdly, this is one of the few existing studies in which coders were allowed to mark multiple labels (e.g., when a person is conveying both fear and sadness in one utterance).

The work by Bruce and Wiebe (1999) in detecting subjective judgments is an interesting, indeed a particularly sophisticated, example of analysis of the results of the annotation to identify the reasons for the disagreement. Bruce and Wiebe had 4 subjects (including 2 participants in the project) assign the tags `subjective` or `objective`, together with a certainty value, to 504 clauses from the Penn Treebank, observing a κ value of 0.599 (using the definition of Davies and Fleiss [1982], our multi-κ). Bruce and Wiebe then applied **correspondence analysis** to the pairwise confusion matrices between the four coders. First, they used the objective/subjective tagging together with the four-point certainty rank to create an eight-point scale ranging from "highly certain subjective" to "highly certain objective". Then, for each coder in a given pair they defined the **profile** for each point on the scale as the vector of the other coder's corresponding judgments. For example, the profile of "highly certain subjective" for coder A is an eight-point vector whose value at each point is the number of items labeled as such by coder B (see Figure 3). Finally, Bruce and Wiebe compared the coders' profiles in order to determine which points on the scale resulted in similar judgment patterns.

Already this allowed Bruce and Wiebe to observe, for instance, that there was much more agreement among their coders on highly certain values than on the highly uncertain ones, and more agreement on marking a clause as "subjective" than "objective". They then applied a combination of techniques for testing the significance of these differences. These techniques allowed them to conclude that although the judges disagreed, strong patterns of 'quasi-symmetry' could be detected, which in turn led

them to explore the use of latent class models (Goodman 1974; Dempster, Laird, and Rubin 1977) to identify what Bruce and Wiebe call the "bias-corrected versions of the judges' original classifications". Bruce and Wiebe apply these techniques to identify the 'latent categories' ('Latent Subjective' and 'Latent Objective') for each item, and propose to use these latent categories as the final classification of the items.

4.3 Marking boundaries and unitizing

Before labeling as just discussed 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 subsection is to assume that the units are linguistic constituents which can be easily identified, such as words for POS tagging, utterances, or noun phrases, and therefore there is no need to check the reliability of this process. We are only 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 other cases however, such as text segmentation and prosodic annotation, 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 two types of annotation, before briefly summarizing Krippendorff's proposals concerning measuring reliability for unitizing.

4.3.1 Segmentation and Topic Marking. The analysis of discourse structure – and especially the identification of discourse segments – is a very important area of research in discourse analysis and computational linguistics, and 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 (Passonneau and Litman 1993; Kowtko, Isard, and Doherty 1992; 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 identifying 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. We focus on segmentation in this section.

Discourse segments are portions of text that constitute a unit either because they are about the same 'topic' or because they have to do with achieving the the same intention (Grosz and Sidner 1986) or performing the same 'dialogue game' (Carletta et al. 1997).[15] The annotation of texts into segments related to the same topic (Hearst 1997; Reynar 1998) is by now a common form of annotation, carried out on a fairly large scale, for example as part of the TREC and then TDT initiatives (Voorhees and Harman 1998; Wayne 2000). These annotation efforts tend to focus on broader analyses such as the division of streams of broadcast news into items about different events, but more fine-grained analyses have also been attempted. For example, Carlson, Marcu, and Okurowski (2003) annotated so-called **discourse units** as the first step of their annotation of discourse structure. An interesting cross between topic-based and rhetor-

15 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 of activity together with all related events and activities'.

ical structure-based analysis is the identification of **argumentative zones** carried out by Teufel, Carletta, and Moens (1999), who segment scientific text according to its role in a scientific text (background, own claims, other people's claims, etc). Carletta et al. (1997) carried out a form of segmentation based on the version of conversational games theory proposed by Sinclair and Coulthard (1975), identifying the boundaries of games and transactions. Intention-based annotations in the Grosz and Sidner sense have also been attempted, although typically on a smaller scale, for example by Passonneau and Litman (1993) and Nakatani et al. (1995), as well as as a part of RDA-style annotations (Moser and Moore 1996; Moser, Moore, and Glendening 1996).

The agreement results in these efforts tend to be on the lower end of the scale proposed by Krippendorff and adopted by Carletta, even for topic-based segmentation. 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 (2001) reports even lower agreement of $K = 0.36$. Teufel, Carletta, and Moens (1999) found higher 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 achieved an agreement 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) managed to achieve very high agreement over unit boundaries (agreement was measured at several times; the initial results were already of $K = 0.87$, and the final result $K = 0.97$). 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 while 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.

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 2001) generally reported lower agreement than the studies whose goal was to identify smaller discourse units is most likely due to the fact that 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. 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 (say, one of the shorter TRAINS dialogues) and we obtain the results in Table 10.

The first case 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

Table 10
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

disagree by 1 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 the 'bulk' of discourse segments, but tend to disagree on their boundaries, 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 4.3.3 below) 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 P_k (Beeferman, Berger, and Lafferty 1999) or WINDOWDIFF (Pevzner and Hearst 2002). Both of these methods aim at assigning partial credit to near misses, and both also specify a metric of disagreement which is additive – that is, the overall disagreement is obtained by adding disagreement over all 'categories' assigned to 'units'. They differ from the methods discussed so far in that the goal of assigning partial credit is achieved by computing pairwise disagreements over the number of boundaries present in a window sliding through the segment; if we allow for these windows to be our 'units', then the way disagreements are computed can be reinterpreted in terms of agreement

coefficients. For instance, we can view P_k as assigning one of two category labels to a window: `same` if both ends of the window are in the same segment, `different` otherwise. The measure P_k now becomes the percentage of windows on which two or more coders agree – that is, observed agreement A_o. Similarly, WINDOWDIFF can be viewed as observed agreement when the category label assigned to each window is the number of boundaries in the window. This highlights the fact that P_k and WINDOWDIFF are simple percentage measures which are not corrected for chance; this can be remedied by reporting K or α instead of observed agreement

Yet another possiblity might be to develop a measure based on the methods used in our own studies of agreement on the antecedents of discourse deixis, which are discourse segments (section 4.4.3 below).

4.3.2 Prosody. Prosodic annotation, like topic marking, crucially involves a step of (prosodic) boundary identification in addition to a step of labeling the units, and measuring agreement on boundaries is as crucial as measuring agreement on the labels. The most important difference from text segmentation is that different types of boundaries exist (Pierrehumbert and Hirschberg 1990). Systematic studies of reliability on both boundary marking and prosodic phrases labeling have been conducted by, among others, Pitrelli, Beckman, and Hirschberg (1994); Syrdal and McGorg (2000); Buhmann et al. (2002). An important difference between these studies is that whereas Syrdal and McGorg (2000) used coders with lots of training, Buhmann et al. (2002) set out to make sure that the annotation could be done reliably by students working at different sites. As in the case of text segmentation, agreement on prosodic segmentation is measured by comparing whether coders classify a word as a boundary or not, and the type of boundary assigned. Syrdal and McGorg (2000) measured agreement between their four expert coders separately on male and female voices, reporting 74% percent agreement on boundaries, for a value of κ of 0.65 for females and 0.62 for males. Buhmann et al. (2002), who tested eight students, do not provide percent agreement, but report κ 'in the range from 0.70 to 0.88' (presumably measured pairwise).

4.3.3 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 proposed by Krippendorff (1995). The full proposal is too complicated to cover here, 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 4); 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 sqare of

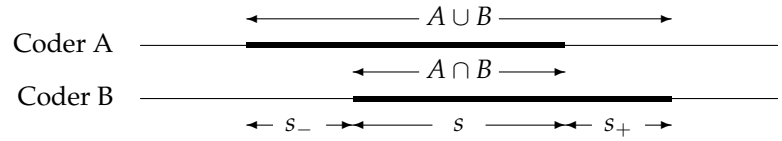

Figure 4
The difference between overlapping units is $\mathbf{d}(A, B) = s_-^2 + s_+^2$ (adapted from Krippendorff 1995, Figure 4, page 61)

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 annotation schemes which require coders to label nested syntactic constituents. Nevertheless, we feel that when the non-overlap assumption holds, testing the reliabilty of unit identification may prove beneficial. Specifically, segmentation (section 4.3.1) can be thought of as a special case of unitizing where all units are of interest, so α_U may serve as a reliability measure for segmentation. 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 since 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, since 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, dialogue d91-3.2).

(3)
```
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 recently carried out (Poesio and Artstein 2005) indicated *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 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. The most reasonable solution to the problem of measuring agreement on anaphoric annotation proposed in the literature is to use as 'labels' the *sets* of mentions of discourse entities, that is, anaphoric / coreference chains (Passonneau 2004). 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. This is because K only offers a dichotomous distinction between agreement and disagreement, whereas practical experience with anaphoric annotation 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 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 to solve 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 Schuetze 1999).

$$\mathbf{d}_J = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad \text{(Jaccard)}$$

$$\mathbf{d}_D = 1 - \frac{2\,|A \cap B|}{|A| + |B|} \qquad \text{(Dice)}$$

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 above (Poesio and Artstein 2005) we used 18 coders to tested α and K under a variety of

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

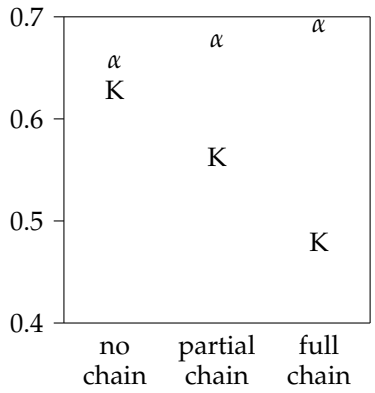

Figure 5
A comparison of the values of α and K for anaphoric annotation

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 like (3) would no longer be considered as disagreements. However, the value of K was drastically reduced, because hardly any coder identified all the mentions of discourse entities (Figure 5).

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 α_κ 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 5, $\alpha = 0.691$. These differences raise an important issue concerning the application of α-like measures for CL tasks: using α is going to make it even more 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 was 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 (4), utterance 7.6 (TRAINS 1991, dialogue d91-2.2).

(4) 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 (Navarretta 2000; Eckert and Strube 2001; 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 its exact boundary: for example, in the case of (4), some marked the text segment as starting 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 since 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 Summarization

Evaluating content selection in summarization is a difficult problem (Radev et al. 2003), for which no single 'gold standard' can exist. (Machine translation faces a similar problem, as do all tasks involving generation.) Even if we only consider the simpler task of comparing summaries obtained by extracting sentences from the original document without any rephrasing, it is extremely unlikely that any two summaries will include ex-

Table 11
Summarization content units produced by two annotators (adapted from Passonneau 2006)

SCU produced by coder 1
Label: Americans asked Saudi officials for help
Weight: 4

Source	Span	
Sum1	1	*Saudi Arabian officials, under American pressure*
Sum2	2	*sought help from Saudi officials*
Sum3	3	*Through the Saudis, the United States asked*
Sum4	4	*U.S. and Saudi Arabian requests*

SCU produced by coder 2
Label: Through the Saudis, the U.S. tried to get cooperation from the Taliban
Weight: 5

Sum1	1	*Saudi Arabian officials, under American pressure,*
	5	*asked Afghan leaders*
Sum3	2	*sought help from Saudi officials,*
	7	*who tried to convince Taleban leaders*
Sum4	3	*Through the Saudis, the United States asked*
Sum5	4	*U.S. and Saudi Arabian requests*
Sum2	6	*U.S. and Saudi officials then attempted*

actly the same sentences – indeed, Lin and Hovy (2003) report that human summarizers agree with their own previous summaries in only about 82% of the cases. The problem is even more complex with more recent evaluation metrics for summarization, which involve identifying the most important 'factoids' contained in reference summaries, and scoring system-produced summaries against these factoids. As one might expect, the coders identifying these common factoids are likely to miss some, just as in the case of anaphoric annotation coders are likely to miss some anaphoric link (Passonneau 2006). Clearly, a measure taking into account partial agreement such as α is needed to measure the agreement between coders producing such a summary.

Indeed, α has been used to measure agreement between coders producing factoid-like evaluation sets out of reference summaries for the DUC competition (Nenkova and Passonneau 2004; Passonneau 2006). Nenkova and Passonneau developed the so-called **pyramid** method for evaluating system-produced summaries, which is based on comparing them with a list of hand-annotated and weighted **summarization content units** (SCU), which are the 'minimal propositions' contained in the text. After expert summarizers have produced reference summaries for each document, coders divide these summaries into SCUs: an SCU is a set of text portions – at most one portion from each reference summary – which express the same factoid. The **weight** of an SCU is the number of summaries which express this factoid. Each SCU is also given a mnemonic **label** which reflects the SCU's content in plain language. Table 11 gives examples of SCUs produced by two coders, expressing related factoids.

The second column in the Table is an indexing of spans of words which allows us to conveniently compare the words associated with each SCU. We see that from some summaries the coders chose identical spans for the SCUs in Table 11 (spans 3 and 4), from others they chose overlapping but non-identical spans (span 1 by coder 1 and spans 1 and 5 by coder 2), and from some summaries only one coder chose a contributing span for this SCU (span 7 by coder 2). We can think of each SCU as a set of spans: coder 2's SCU is the set $\{1, 2, 3, 4, 5, 6, 7\}$, while coder 1's SCU is the set $\{1, 2, 3, 4\}$ (coder 1 included spans 5 and 7 in a separate SCU, and span 6 in yet another SCU).

SCU identification can now be seen as dividing a text into sets of spans; this is similar to anaphoric annotation, where markables are divided into sets which form coreference chains. To measure the reliability of SCU identification, Nenkova and Passonneau (2004) use Krippendorff's α with the distance metric of Passonneau (2004), while Passonneau (2006) uses the newer MASI distance metric (section 4.4.1).

4.6 Word Senses

Wordsense 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 wordsense 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 wordsense 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 subjects 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 6 different naive subjects. 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 wide variety of values when measured per word – ranging from 0.007 for adjective *correct* to 0.92 for 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

Table 12
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*.

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 (Buitelaar 1998; Bruce and Wiebe 1998; Véronis 1998; Palmer, Dang, and Fellbaum 2007). Hierarchical tagsets were also developed, such as HECTOR (Atkins 1993) or, indeed, WordNet itself (where senses are related by hyponymy links). In the case of Buitelaar (1998) and Palmer, Dang, and Fellbaum (2007), the 'supersenses' were identified by hand, whereas Bruce and Wiebe (1998) and Véronis (1998) used clustering methods such as those from Bruce and Wiebe (1999) to collapse some of the initial sense distinctions. 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 12).

Palmer, Dang, and Fellbaum 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 the value of K from 0.874 to 0.898. Using a related technique, Véronis (1998) found that agreement on noun wordsense tagging went up from a K of around 0.45 to a K of 0.86.[16]

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 method for computing K with 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

16 We are not aware of any annotation effort attempting to use a tagset based on Buitelaar's CORELEX, a reconstruction of the WordNet repertoire of noun wordsenses according to Pustejovsky's Generative Lexicon theory (Buitelaar 1998).

on the leaves. For example, with a tagset like the one in Table 12, 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 Dan 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 ad B are identical, 1 when the tags do not overlap, and somewhere inbetween 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 (Dan 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 wordsense 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):

(5) 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 (5) 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 {WN3,**LABEL**}, which would give in 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 wordsense annotation was developed and tested by Passonneau, Habash, and Rambow (2006). The approach of Passonneau, Habash, and Rambow 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). The problem with this approach is that in practice, coders very seldom assign more than one label to units (Véronis 1998; Poesio and Artstein 2005).

5. Other issues

5.1 Missing Data

An assumption that underlies all the coefficients that we have discussed is that all the coders classify all the items in the reliability sample. In practice, however, this is not always the case, either because of practical limitations on the experimental setup or because some of the coders fail to classify certain items, for whatever reason. When data points are missing, the coefficients need to be adjusted to minimize the loss.

If there are only two coders then missing data implies the existence of items with at most one judgment, and since such singular judgments cannot be compared with anything, the only remedy is to remove these items from the sample. We can thus only deal with missing data when the number of coders is three or greater.

We will not attempt to deal with missing data in coefficients that model chance using individual coder marginals (multi-κ, α_κ). While this is possible in principle, and we have done so in a previous version of this article, the formulas become very complicated and their usefulness is quite dubious, because the values of single- and individual-distribution coefficients converge with multiple coders (section 3.1). We will therefore only deal with single-distribution coefficients in this section (multi-π, α). Note also that no adjustment is needed if different items are classified by different coders, as long as the number of coders per item is constant, because these coefficients treat coders as interchangeable. If, however, the number of judgments per item is not constant, then the coefficients need to be adjusted.

One solution to the missing data problem is to eliminate certain data points in order to achieve a data set were all coders classify all items. This is probably the best practice when the total data loss would be small. In a recent annotation experiment (Poesio and Artstein 2005) we had a total of 151 items, and data points were missing for three of them; eliminating these items from the analysis provided a quick solution to the missing data problem. Another example is Fleiss (1971), which reports a psychiatric study where each patient (item) was diagnosed by between 6 and 10 psychiatrists (coders), and

which reduced the number of coders per patient to a constant (six) by random dropping of diagnoses. There is no indication of how much data were lost in this pruning.

An alternative to dropping additional data is to redefine the observed and expected agreement and disagreement so as to minimize the skewing of the coefficient values. We will present two ways of doing this, either giving equal weight to each *judgment* or to each *item*. The method which gives an equal weight to each judgment is advocated by Krippendorff (2004a, pages 230–232) for use with α; it is in line with the origins of α as a measure of variance, since the standard method of computing the F stastisic in an analysis of variance involves giving equal weight to each data point, not to each level. Krippendorff's justification for giving an equal weight to each judgment is that the total number of judgments is the best estimate of the actual distribution of items; however, such an estimate may be skewed if judgements are missing for a systematic reason (for example, if judgments are missing for items of one particular category). For this reason we also provide a way to adjust the coefficients in order to give an equal weight to each item.

We start by calculating a normalized figure for observed agreement and disagreement per item. Let $\mathbf{n}_i$ stand for the number of judgments available for a particular item i; the total number of judgment pairs for item i is therefore $\binom{\mathbf{n}_i}{2} = \mathbf{n}_i(\mathbf{n}_i - 1)/2$. Normalized agreement is the total number of agreeing judgment pairs and normalized disagreement is the total distance between the judgment pairs, both divided by the total number of judgment pairs.

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

$$\mathrm{disagr}_i = \frac{1}{\mathbf{n}_i(\mathbf{n}_i - 1)} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \mathbf{n}_{ik_j} \mathbf{n}_{ik_l} \mathbf{d}_{k_j k_l}$$

If an item i receives only one judgment, then agr_i and disagr_i turn out to be $\frac{0}{0}$, or undefined; indeed in such a case one cannot talk of agreeing or disagreeing pairs at all. The overall observed (dis)agreement is calculated only on the items for which such a value exists. Let I' be the set of items for which there are two or more judgments available, let $\mathbf{i}'$ be the cardinality of this set, and let N' be the total number of judgments available for this set. Giving an equal weight to each judgment, observed (dis)agreement is the mean of the normalized (dis)agreement values of the individual items, weighted by the number of judgments per item.

$$\mathrm{A}_\mathrm{o}^\pi = \frac{1}{N'} \sum_{i \in I'} \mathbf{n}_i \mathrm{agr}_i \qquad \mathrm{D}_\mathrm{o}^\alpha = \frac{1}{N'} \sum_{i \in I'} \mathbf{n}_i \mathrm{disagr}_i$$

Giving an equal weight to each item, observed (dis)agreement is the unweighted mean of the (dis)agreement values of the individual items.

$$\mathrm{A}_\mathrm{o}^\pi = \frac{1}{\mathbf{i}'} \sum_{i \in I'} \mathrm{agr}_i \qquad \mathrm{D}_\mathrm{o}^\alpha = \frac{1}{\mathbf{i}'} \sum_{i \in I'} \mathrm{disagr}_i$$

The formulas for expected agreement and disagreement already give equal weight to each judgment, so the only modification necessary is to remove all singulary judgments before computation. Let $\mathbf{n}'_k$ be the total number of judgments of category k that

come from items for which there are at least two judgments in total. Expected agreement and disagreement then take the following shape.

$$A_e^{\pi} = \frac{1}{(N')^2} \sum_{k \in K} (\mathbf{n}'_k)^2 \qquad D_e^{\alpha} = \frac{1}{N'(N'-1)} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \mathbf{n}'_{k_j} \mathbf{n}'_{k_l} \mathbf{d}_{k_j k_l}$$

If we want to give equal weight to each item, we have to weight each individual judgment by the inverse of the number of judgments in the item it comes from.

$$A_e^{\pi} = \frac{1}{(\mathbf{i}')^2} \sum_{k \in K} \left(\sum_{i \in I'} \frac{\mathbf{n}_{ik}}{\mathbf{n}_i} \right)^2 \qquad D_e^{\alpha} = \frac{N'}{N'-1} \frac{1}{(\mathbf{i}')^2} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \left(\sum_{i \in I'} \frac{\mathbf{n}_{ik_j}}{\mathbf{n}_i} \right) \left(\sum_{i \in I'} \frac{\mathbf{n}_{ik_l}}{\mathbf{n}_i} \right) \mathbf{d}_{k_j k_l}$$

5.2 Error models

One of the problems with reliability coefficients is that the value they give is not easily interpretable as a parameter of the population of items or of the coding process. It would be desirable to replace the notion 'amount of agreement' with something easier to understand, like error rate. It is not possible to determine the error rate without knowledge of the true category labels; however, an explicit model of annotator error can estimate error rates from the annotation data. In this section we will look at an explicit model of annotator error proposed by Aickin (1990), and show how it can be used to estimate error rates from agreement coefficients.

Aickin's alpha. Aickin (1990) offers a way to measure agreement between two coders using an explicit model with a parameter called 'alpha'; we will designate it with the symbol **a** in order to distinguish it from Krippendorff's α. The model is based on four assumptions, the first three of which are as follows.

1. The items are made up of two populations, one which is easy to classify and one which is hard to classify.
2. The coders always agree on the classification of the easy items.
3. The coders classify hard items at random.

These assumptions give rise to an attractive conceptualization of the reliability problem. From our experience, items really differ in difficulty, and while it is clearly an oversimplification to make a dichotomous distinction between items that are classified at 100% accuracy and others that are classified at chance level, this does capture a basic intuition about an important source of disagreement in annotation.

The population parameter **a** is the proportion of items which are easy to classify. Since agreement is maximal for the easy items and at chance for the hard items, observed agreement and disagreement turn out to be functions of just the propotion **a** and the expected agreement or disagreement on the population of hard items (denoted by A_e^{hard} and D_e^{hard}).

$$A_o = \mathbf{a} + (1 - \mathbf{a}) A_e^{hard} \qquad D_o = (1 - \mathbf{a}) D_e^{hard}$$

The above formulas can be transformed into a form which is similar to that of the familiar coefficients.

$$\mathbf{a} = \frac{A_o - A_e^{hard}}{1 - A_e^{hard}} \qquad \mathbf{a} = 1 - \frac{D_o}{D_e^{hard}}$$

The difference between this and kappa-like coefficients is that chance agreement is calculated only on the items which are assumed to be randomly classified.

If we add to this model the assumption that $A_e = A_e^{hard}$ or $D_e = D_e^{hard}$, we find that **a** is identical to π, κ, or α, depending on how exactly expected agreement is calculated. A natural interpretation of the assumption $A_e = A_e^{hard}$ is that the distribution of judgments for the hard items is determined by the distribution of the easy items, on which there is agreement (for κ this also entails identical coder marginals). We thus have a new interpretation of the agreement coefficients, namely the proportion of items which are easy to classify under an error model consisting of the first three premises of Aickin (1990) plus the assumption that $A_e = A_e^{hard}$ or $D_e = D_e^{hard}$ (see also Krippendorff 2004a, page 227).

Aickin's fourth assumption is different from our last assumption, and is intended to achieve a property called **constant predictive probability**: if the coders agree on the classification of an item, the probability that this item is easy is the same irrespective of the category to which it was assigned (Aickin only considers annotations performed by two coders, not more). The following assumption achieves this property.

4. constant predictive probability: Easy items are distributed among those category pairs that denote agreement in proportion to the distribution of the hard items among these category pairs.

With **k** category labels, the model has $2\mathbf{k} - 1$ parameters: two probability distributions of the hard items (one for each coder, each characterized by $\mathbf{k} - 1$ parameters), plus the proportion of easy items **a**.

Aickin uses maximum-likelihood estimation to estimate the parameters of the model, and reports the results of simulations which were conducted with identical distributions for the two coders. The simulations show that the value of **a**, estimated by maximum likelihood, tends to be higher than that of κ, except when the distributions are uniform, in which case the two coefficients yield similar values. This is no surprise, as it can be shown analytically that $\kappa \leq \mathbf{a}$ in any data set generated by the model for coders with identical distributions, with the limiting case obtaining only when the coders' distributions are uniform (we omit the proof for lack of space). The reason for this is the fact that Aickin's assumption 4 introduces a *nonlinear* relation between the distribution of the easy items and that of the hard items.

We feel that Aickin's assumption 4 does not have a natural interpretation: it seems to imply that the distribution of judgments for the hard items is determined by the *square roots* of the proportions of the easy items, suggesting that the coders are somehow aware of the method of calculating agreement by looking at their joint decisions. We also have doubts about the constant predictive probability property, which is the driving force behind assumption 4. The chance of spurious classification of an item into a common category is higher than the chance of spurious classification into a rare category, and therefore agreement on the classification of an item into a common category should

indeed be less indicative that this agreement is genuine. This intuition, which we feel is valid, is contrary to the constant predictive probability hypothesis.

Finally, we conjecture that simulations of an error model which replaces assumption 4 with the assumption that $A_o = A_o^{hard}$ would find the maximum-likelihood estimator for **a** very similar to the values of π and κ even when the distribution of items among categories is not uniform. We expect this at least for coding data with a substantial amount of agreement. The error model cannot generate data with systematic disagreement, and will therefore be a poor model for such data; with systematic disagreement, π and κ can dip below zero, while zero is by definition the lower bound for **a**.

Randomly spread error. In the model of Aickin (1990), items are either easy or hard to classify, and arbitrary (chance) judgments are given to all and only the hard items. We can also construct a different error model, in which arbitrary judgments are spread evenly over all items. For each judgment there will be a single, invariant probability p of making a non-arbitrary (correct) classification, and a probability of $1 - p$ of making an arbitrary decision according to the true distribution of items among categories $P(k_1), \ldots, P(k_{\mathbf{k}})$ (the arbitrary decision can also turn out to be correct, by chance). We can think of the coding process as going in two steps: first the true category k of an item is determined, and then each coder classifies the item as k with probability $p + (1 - p)P(k)$, and as any other category $k' \neq k$ with probability $(1 - p)P(k')$. Under this error model, the expected value of π is p^2 – that is, it depends only on the proportion of non-arbitrary judgments.

$$
\begin{aligned}
E(A_o) &= \sum_{k \in K} P(k) \Big((p + (1-p)P(k))^2 + \sum_{k' \neq k} ((1-p)P(k'))^2 \Big) \\
&= \sum_{k \in K} P(k) \Big(p^2 + 2p(1-p)P(k) + (1-p)^2 \sum_{k \in K} (P(k))^2 \Big) \\
&= p^2 + 2p(1-p)A_e^\pi + (1-p)^2 A_e^\pi \\
&= p^2 + (1-p^2)A_e^\pi \\
E(\pi) &= \frac{A_o - A_e^\pi}{1 - A_e^\pi} = \frac{p^2 + (1-p^2)A_e^\pi - A_e^\pi}{1 - A_e^\pi} = \frac{p^2 - p^2 A_e^\pi}{1 - A_e^\pi} = p^2
\end{aligned}
$$

We can compare this to the result from the discussion of Aickin's alpha: when arbitrary judgments are distributed in proportion to the actual distribution of items among categories and all the arbitrary judgments concentrate on the same items, then π is equal to the proportion of non-arbitrary judgments; if arbitrary judgments are spread evenly among items, then π is equal to the *square* of the proportion of non-arbitrary judgments. In practice we expect the arbitrary judgments to lie somewhere between these extremes, that is to be somewhat arbitrary, with a tendency to concentrate on the more difficult items. We thus expect the proportion of non-arbitrary judgments to be somewhere between π and $\sqrt{\pi}$. The proportion of correct judgments will be somewhat higher, since some arbitrary judgments will also be correct.

6. Conclusions

We conclude this fairly long discussion by summarizing what in our view are the main points emerging from ten years of CL experience with chance-corrected coefficients of agreement. These points can be grouped under three main headings: methodology, choice of coefficients, and interpretation of coefficients.

6.1 Methodology

One clear result of our survey has already been announced at the beginning of section 4: still too few studies of the reliability of a coding scheme in CL apply a methodology as rigorous as that envisaged by Krippendorff. All too often, agreement studies are just a race to get a high score. It is true that the methodology adopted for large annotation efforts has greatly improved: one need only compare the central role played by reliability testing in the case of the Penn Discourse Treebank (Miltsakaki et al. 2004) or OntoNotes (Hovy et al. 2006) with the absence of any tests in the case of the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) or the British National Corpus (Leech, Garside, and Bryant 1994). But even in the case of such large annotation efforts, only percent agreement gets measured. There are a number of reasons for this. One is that annotation efforts tend to be carried out by engineers (who often do not have the time for a rigorous test) or by linguists (many of whom do not believe that untrained subjects can make sound linguistic judgments). But the difficulty in interpreting the results also plays a role: many researchers do not see the point in carrying out a reliability study if then they can't interpret its results. Still, we find this status of affairs rather unsatisfactory, as in our experience coefficients of agreement, together with the other information discussed below, do provide a better indication of the quality of the resulting annotation than simple percent agreement.

One area in which a motivated difference may be emerging between Content Analysis practice and CL methodology is in the role of experts. The main concern in Content Analysis is to ensure that the results of a given study are reproducible; to guarantee this, one must make sure that the coding on which these results are based can be reliably reproduced by individuals whose only training is provided by the coding scheme. One of the main purposes of reliability testing in CL, on the other hand, is to test schemes used for resource creation – that is, the 'result' will be an annotated corpus, not some scientific claim. Now, most annotation tasks of interest to CL require judgments which are too complex to be drawn by naive coders, so in practice professionals rather than naive coders are employed for serious resource creation efforts whenever financial resources allow. This can be achieved either by using professionals, as advocated by Kilgarriff (1999), or through an intensive training, as done in OntoNotes (Hovy et al. 2006) or by Carlson, Marcu, and Okurowski (2003). We think this is not problematic so long as precautions are taken to ensure that the resource will be consistent – that is, to ensure that the people who will actually do the job are consistent and agree with each other. So long as the people who actually do the job are experts, we feel it is adequate to draw the coders for the reliability test from this kind of population. This practice will certainly not offer the same cast-iron guarantee as using naive coders who only follow written instructions, and the resulting corpora cannot be used to make claims about spontaneous linguistic judgments, but in practice the only alternative would be to limit corpus annotation in CL to the kind of "oversimplified or superficial but reliable text analyses" that quite rightly Krippendorff finds to be of as limited usefulness as

"fascinating interpretations that nobody can replicate" (Krippendorff 2004a, pages 213–214).

6.2 Choosing a coefficient

Up until the papers by Passonneau (2004) and Di Eugenio and Glass (2004), K was viewed for all intents and purposes as the only available option for measuring reliability. One of the goals of this paper is to further the process of reconsideration of the available options.

As far as we know, the debate concerning the 'horizontal' dimension of the coefficient cube – annotator bias – has been limited to the exchange between Di Eugenio and Glass (2004), in favor of Cohen's κ, and Craggs and McGee Wood (2005) (and Krippendorff), in favor of K and α. There is an overwhelming consensus in CL practice: K and α are used in the vast majority of the studies we reported. We also incline towards the view that K and α are more appropriate, as they abstract away from the bias of specific coders, as well as not suffering from the unpleasant properties noted by Di Eugenio and Glass. 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), and we believe that increasing the number of annotators is the best strategy, also to increase the variability of data.

One of the main goals of this article has been to bring to the attention of the CL community the fact that in many cases an argument can be made for using weighted coefficients, and that the choice between weighted and unweighted measures is of greater import. We think there are at least two types of coding schemes in which partial agreement measures may be considered:

- Coding schemes with hierarchical tagsets
- Coding schemes with set-valued interpretations (anaphora, summarization)

We discussed various examples of both, and argued that at least in the second case, weighted coefficients are almost unavoidable. The problem is that the results obtained with these measures are not easy to interpret. Our suggestion would therefore be as follows.

- Use clearly disjoint labels and a binary distance function when possible (that is, K);
- Use weighted measures when the task demands it, but then do not expect to be able to interpret the value thus obtained using scales such as those proposed by Krippendorff or Landis and Koch.

6.3 Interpreting the values

We perceive the lack of consensus on how to interpret the value of the coefficient of agreement as the most serious problem with current practice in reliability testing, and one of the main reasons for the reluctance of many in CL to embark in reliability studies.

We already said that Krippendorff's position is quite clearly that a value of 0.8 is the absolute minimum for any serious claims to be supported by the data. As far as resource creation is concerned, our own experience is more consistent with Krippendorff and

Neuendorf's than with that of Landis and Koch: 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).

However, it is doubtful that a single cutoff point is appropriate for all purposes. Even the lower 0.67 level has often proved impossible to achieve in CL research, particularly on discourse, except via substantial training (see, e.g., Hearst 1997; Poesio and Vieira 1998); often, substantial agreement among coders results in values of K or α around the 0.7 level replicated across studies. Provided that significance is reached, we feel that this level of agreement may be all that one may hope to achieve for certain types of judgments, and we agree therefore with Craggs and McGee Wood (2005) that insisting that the magic 0.8 threshold be reached is unhealthy; on this, see also Krippendorff's remarks about losing validity to reach reliability. (We especially hope this paper won't result in readers viewing weighted coefficients as a particularly nifty trick to raise their K score!)

Unfortunately, weighted coefficients, while arguably more appropriate for many annotation tasks, as we have seen, make the issue of deciding when the value of a coefficient indicate sufficient agreement even more complicated. With weighted measures, the value of the coefficient greatly depends on the distance metric chosen, as we saw at the end of section 4.4.

This being the situation at the moment, however, we feel that simply reporting the value of a chance-corrected coefficient of agreement is not informative enough. Given that coefficients such as K or α do not have a clear interpretation, and given also the distorting effects of skewed distributions, simply reporting the value of K is not enough in order to understand what the results actually mean. On this point we agree with Di Eugenio and Glass (2004), but we feel that their solution of reporting the values of more than one coefficient of agreement is not the right solution. Instead, researchers should clearly report the methodology that was followed to collect the reliability data (number of coders, whether they coded independently, whether they relied exclusively on an annotation manual). The study should also indicate whether agreement was statistically significant, and provide the 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).

One approach that in our opinion may offer a way out of the problem of interpreting the results is **Latent class analysis** (Uebersax 1988; Uebersax and Grove 1990), a term used to indicate the application of **latent class modeling** techniques to nominal data (the term **latent trait analysis** is used for ordinal categories). Latent class modeling techniques, of which perhaps the best known example is the familiar EM algorithm (Goodman 1974; Dempster, Laird, and Rubin 1977), were developed to deal with classification tasks in which the pattern of results (say, the POS tags assigned to words) derives from the membership of the items that have to be classified to an (unknown) number of (unknown) categories (the latent classes). Uebersax and colleagues show how latent class analysis methods can be used to analyze agreement, and argue that such methods solve many of the problems they find with kappa-like coefficients of agreement (Uebersax 1988; Uebersax and Grove 1990). A full discussion of these methods falls outside the scope of this article, but we think they are worth mentioning, especially as a way of addressing the problem of interpreting the value of K.

Uebersax and colleagues view the coding process as an instance of the noisy channel model. An item, whose real class is unobservable (latent), is assigned a category label by

coders according to the probability distribution associated with the latent class; the only observable results are the annotators' category labels. Using standard techniques in CL, we can however estimate the probability $P(\text{label}_i = l|\text{latent-class}_i = c)$ using maximum likelihood estimation to get initial estimates from the observed frequency counts, and then the EM algorithm to get more accurate estimates. For instance, returning to our first example in Table 1, we could estimate $P(\text{label}_i = \text{STAT}|\text{latent-class}_i = \text{STATEMENT})$ and similarly for IREQ. We could then use these probabilities to evaluate the reliability of our coding procedure: for example, we could decide that our coding procedure is sufficiently reliable if such probabilities are sufficiently high (say, higher than 95%) or, alternatively, the probability of error is sufficiently low.

We find this approach extremely promising, but to our knowledge it has not yet been used as a way of evaluating agreement on annotation in CL. (Latent class modeling has been used by Bruce and Wiebe (1998) in order to identify the number of latent classes that seemed to underly the behavior of their coders, but not to analyze agreement.)

Appendix A. Bias and variance with multiple coders

In section 3.1 we briefly noted that the difference between π and κ drops as the number of coders increases, because this difference is the overall variance of the different categories divided by the number of annotators. Here we give the formal proof. We start by taking the formulas for expected agreement from section 2.5 and putting them into a form that is more useful for comparison with one another.

$$\begin{aligned}
A_e^\pi &= \sum_{k\in K} \hat{P}(k)^2 = \sum_{k\in K} \left(\frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \hat{P}(k|c_m) \right)^2 \\
&= \sum_{k\in K} \frac{1}{\mathbf{c}^2} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{P}(k|c_m)\hat{P}(k|c_n) \\
A_e^\kappa &= \sum_{k\in K} \frac{1}{\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \hat{P}(k|c_m)\hat{P}(k|c_n) \\
&= \sum_{k\in K} \frac{1}{\mathbf{c}(\mathbf{c}-1)} \left(\sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{P}(k|c_m)\hat{P}(k|c_n) - \sum_{m=1}^{\mathbf{c}} \hat{P}(k|c_m)^2 \right)
\end{aligned}$$

The overall annotator bias B is the difference between the expected agreement according to π and the expected agreement according to κ.

$$\begin{aligned}
B &= A_e^\pi - A_e^\kappa \\
&= \frac{1}{\mathbf{c}-1} \sum_{k\in K} \frac{1}{\mathbf{c}^2} \left(\mathbf{c} \sum_{m=1}^{\mathbf{c}} \hat{P}(k|c_m)^2 - \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{P}(k|c_m)\hat{P}(k|c_n) \right)
\end{aligned}$$

We now calculate the mean μ and variance σ^2 of $\hat{P}(k|c)$, taking c to be a random variable with equal probabilities for all of the coders: $\hat{P}(c) = \frac{1}{\mathbf{c}}$ for all coders $c \in C$.

$$\mu_{\hat{P}(k|c)} = \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \hat{P}(k|c_m)$$

$$\begin{aligned}
\sigma^2_{\hat{\mathrm{P}}(k|c)} &= \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} (\hat{\mathrm{P}}(k|c_m) - \mu_{\hat{\mathrm{P}}(k|c)})^2 \\
&= \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k|c_m)^2 - 2\mu_{\hat{\mathrm{P}}(k|c)} \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k|c_m) + \mu^2_{\hat{\mathrm{P}}(k|c)} \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} 1 \\
&= \left(\frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k|c_m)^2 \right) - \mu^2_{\hat{\mathrm{P}}(k|c)} \\
&= \frac{1}{\mathbf{c}^2} \left(\mathbf{c} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k|c_m)^2 - \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{\mathrm{P}}(k|c_m)\hat{\mathrm{P}}(k|c_n) \right)
\end{aligned}$$

The annotator bias B is thus the sum of the variances of $\hat{\mathrm{P}}(k|c)$ for all categories $k \in K$, divided by the number of coders less one.

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

Since the variance does not increase in proportion to the number of coders, we find that the more coders we have, the lower the annotator bias; at the limit, κ approaches π as the number of coders approaches infinity.

Appendix B. Bias of weighted measures

We have shown in appendix A that the variance of the individual coders' distibutions of items to categories is a useful measure for the annotator bias in a set of coding data, and that it correlates with the difference between π and κ. This measure of variance is less useful when the coding data are judged according to a weighted measure, because the discrepancies between the individual coders also have varying magnitudes. A measure of annotator bias for such coding data should therefore take the weights into account. Since the expected disagreement already considers the weights, we define the annotator bias B in an analogous way to our definition in appendix A, namely as the difference between the expected disagreement according to the single-distribution measure α_b and the expected disagreement according to the individual-distribution measure α_κ.

$$\mathrm{B} = \mathrm{D}_\mathrm{e}^{\alpha_b} - \mathrm{D}_\mathrm{e}^{\alpha_\kappa}$$

We first put the expected disagreements according to α_b and α_κ (sections 2.6 and 2.7 respectively) into forms that are more useful for the comparison.

$$\mathrm{D}_\mathrm{e}^{\alpha_b} = \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \hat{\mathrm{P}}(k_j)\hat{\mathrm{P}}(k_l)\mathbf{d}_{k_j k_l} = \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \frac{1}{\mathbf{c}^2} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_n)\mathbf{d}_{k_j k_l}$$

$$\mathrm{D}_\mathrm{e}^{\alpha_\kappa} = \frac{1}{\binom{\mathbf{c}}{2}} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_n)\mathbf{d}_{k_j k_l}$$

$$= \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \frac{1}{\mathbf{c}(\mathbf{c}-1)} \left(\sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_n) - \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_m) \right) \mathbf{d}_{k_j k_l}$$

Now we calculate the annotator bias as the difference between the above measures.

$$\begin{aligned}
\mathrm{B} &= \mathrm{D}_{\mathrm{e}}^{\alpha_b} - \mathrm{D}_{\mathrm{e}}^{\alpha_\kappa} \\
&= \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \left((\frac{1}{\mathbf{c}^2} - \frac{1}{\mathbf{c}(\mathbf{c}-1)}) \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_n) \right. \\
&\qquad \left. + \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_m) \right) \mathbf{d}_{k_j k_l} \\
&= \frac{1}{\mathbf{c}-1} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \frac{1}{\mathbf{c}^2} \left(\mathbf{c} \sum_{m=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_m) - \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \hat{\mathrm{P}}(k_j|c_m)\hat{\mathrm{P}}(k_l|c_n) \right) \mathbf{d}_{k_j k_l}
\end{aligned}$$

Unlike the case for unweighted measures, this measure of annotator bias does not correspond to the sum of the variances of a single random variable. But the bias still drops in proportion to an increase in the number of coders: the sums inside the parentheses grow in proportion to $\mathbf{c}^2$, and therefore the overall annotator bias B grows in proportion to $1/(\mathbf{c}-1)$.

Acknowledgments

This work was in part supported by EPSRC grant GR/S76434/01, ARRAU. We wish to thank four anonymous reviewers and 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

Aickin, Mikel. 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, 46(2):293–302.

Allen, James and Mark Core. 1997. DAMSL: Dialogue act markup in several layers. Draft contribution for the Discourse Resource Initiative.

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, Germany.

Atkins, S. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX 93*, Budapest.

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. *Biomet-*

rics, 45(1):269–287.

Brants, T. and O. Plaehn. 2000. Interactive corpus annotation. In *Proc. 2nd LREC*, Athens.

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 [F.] and Janyce [M.] Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of EMNLP*, pages 53–60, Granada, Spain.

Bruce, Rebecca F. and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

Buhmann, J., J. Caspers, V. van Heuven, H. Hoekstra, J. P. Martens, and M. Swerts. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Ccorpus. In *Proc. of LREC*, pages 779–785, Las Palmas.

Buitelaar, Paul. 1998. *CoreLex : Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Bunt, H. C. 2000. Dynamic interpretation and dialogue theory. In M. M. Taylor, F. Néel, and D. G. Bouwhuis, editors, *The Structure of Multimodal Dialogue II*. John Benjamins, Amsterdam, pages 139–166.

Bunt, H. C. 2005. A framework for dialogue act specification. In *Proc. Joint ISO-ACL Workshop on the Representation and Annotation of Semantic Information*, Tilburg.

Byron, D. 2002. Resolving pronominal references to abstract entities. In *Proc. of the ACL*, pages 80–87.

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, L., D. Marcu, and M. E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer, pages 85–112.

Chinchor, N. A. 1997. MUC-7 named entity task definition. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*. Available at `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html`.

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.

Cimiano, P. and S. Handschuh. 2003. Ontology-based linguistic annotation. In *Proc. of the ACL Workshop on Linguistic Annotation*, pages 14–21, Sapporo, Japan.

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, M. G. and J. F. Allen. 1997. Coding dialogs with the DAMSL scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, AAAI, Boston, MA.

Craggs, Richard and Mary McGee Wood. 2004. A two-dimensional annotation scheme for emotion in dialogue. In *Proc. of AAAI Spring Symmposium on Exploring Attitude and Affect in Text*, Stanford.

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.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Di Eugenio, B., P. W. Jordan, J. D. Moore, and R. H. Thomason. 1998. An empirical investigation of collaborative dialogues. In *Proc. of the 36th ACL*, Montreal.

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.

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

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassell, and R. Weischedel. 2000. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proc. of LREC*.

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

Doran, C., J. Aberdeen, L. Damianos, and L. Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proc. of 2nd SIGDIAL workshop*.

Eckert, M. and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

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.

Fellbaum, Christiane, Joachim Grabowski, and Shari Landes. 1997. Analysis of a hand-tagging task. In *Proceedings of ANLP Workshop on Tagging Text with Lexical Semantics*, pages 34–40, Washington, DC.

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.

Geertzen, J. and H. Bunt. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proc. of the 7th SIGDIAL*, Sydney, Australia.

Goodman, Leo A. 1974. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – a modified latent structure approach. *American Journal of Sociology*, 79(5):1179–1259.

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.

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

Handschuh, S. 2005. *Creating Ontology-based Metadata by Annotation for the Semantic Web*. Ph.D. thesis, Universität Karlsruhe.

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, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proc. HLT-NAACL*.

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, S., A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. 1995. Dialogue acts in VERBMOBIL. VERBMOBIL-Report 65, Universität Hamburg, DFKI GmbH, and Universität Erlangen.

Jurafsky, D., E. Shriberg, and D. Biasca. 1997. Switchboard-DAMSL labeling project coder's manual. Technical Report 97-02, University of Colorado at Boulder, Institute for Cognitive Science, Boulder, Colorado. Available from http://www.colorado.edu/ling/jurafsky/manual.august1.html.

Kilgarriff, A. 1999. 95% replicability for manual word sense tagging. In *Proc. of the EACL (Poster session)*.

Kowtko, J. C., S. D. Isard, and G. M. Doherty. 1992. Conversational games within dialogue. Research Paper HCRC/RP-31, Human Communication Research Centre.

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

Krippendorff, Klaus. 1970a. Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, 2:139–150.

Krippendorff, Klaus. 1970b. 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. 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, G., R. Garside, and M. Bryant. 1994. Claws4: The tagging of the british national corpus. In *Proc. 15th COLING*, pages 622–628, Kyoto.

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

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Edmonton.

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

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, M. P., B. Santorini, and M. A. Marcinkiewicz. 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.

Melamed, I. Dan and Philip Resnik. 2000. Tagger evaluation given hierarchical tagsets. *Computers and the Humanities*, 34(1–2):79–84.

Mieskes, M. and M. Strube. 2006. Part-of-speech tagging of transcribed speech. In *Proc. of 5th LREC*, pages 935–938, Genoa, Italy.

Mihalcea, R., T. Chklovski, and A. Kilgarriff. 2004. The SENSEVAL-3 english lexical sample task. In *Proc. SENSEVAL-3*, Barcelona.

Miltsakaki, E., R. Prasad, A. Joshi, and B. Webber. 2004. Annotating discourse connectives and their arguments. In *Proc. NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston.

Moser, M. and J. D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

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

Nakatani, C. H., B. J. Grosz, D. D. Ahn, and J. Hirschberg. 1995. Instructions for annotating discourses. Technical Report TR-25-95, Harvard University Center for Research in Computing Technology.

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

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

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, R. J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. of 5th LREC*, Genoa.

Passonneau, R. J., N. Habash, and O. Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proc. of 5th LREC*, Genoa.

Passonneau, R. J. and D. J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1).

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

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.

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

Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in english. In J. Morgan P. Cohen and M. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA, pages 271–312.

Pitrelli, J., M. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labelling reliability in the TOBI framework. In *Proc. of 3rd ICSLP*, volume 2, pages 123–

126, Yokohama.

Poesio, M. and N. N. Modjeska. 2005. Focus, activation, and this-noun phrases: An empirical study. In A. Branco, R. McEnery, and R. Mitkov, editors, *Anaphora Processing*. John Benjamins, pages 429–442.

Poesio, M., A. Patel, and B. 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*, 4:229–257. Special Issue on Generation and Dialogue.

Poesio, Massimo. 2004a. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, 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, Association for Computational Linguistics, Cambridge, Massachusetts.

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, Association for Computational Linguistics, Ann Arbor, Michigan.

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

Popescu-Belis, A. 2005. Dialogue acts: One or more dimensions? Working Paper 62, ISSCO, 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.

Radev, Dragomir R., Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of 41st Annual Meeting of the ACL*, pages 375–382, Sapporo.

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

Reinhart, T. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1). Also distributed by Indiana University Linguistics Club.

Reynar, J. C. 1998. *Topic Segmentation: Algorithms and Applications*. Phd, University of Pennsylvania, IRCS, Philadelphia, PA.

Ries, K. 2001. Segmenting conversations by topic, initiative and style. In *Proc. ACM SIGIR Workshop on Information Retrieval Techniques for Speech Applications*, New Orleans.

Rietveld, T. and R. van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behavior*. Mouton de Gruyter.

Rosenberg, A. and E. Binkowski. 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proc. of NAACL*, volume Short papers.

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

Sekine, S., K. Sudo, and C. Nobata. 2002. Extended named entity hierarchy. In *Proc. of LREC*.

Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. of 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.

Sinclair, J. M. and R. M Coulthard. 1975. *Towards an analysis of discourse: the English used by teachers and pupils*. Oxford University Press.

Stent, A. 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, University of Rochester, Department of Computer Science, Rochester, NY.

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

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van-Ess-Dykema, and M. Meteer. 1997. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371.

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

Syrdal, A. and J. McGorg. 2000. Inter-transcriber reliability of ToBi prosodic labelling. In *Proc. of 6th ICSLP*, volume 3, pages 235–238, Bejing.

Tateisi, T., Y. Ohta, N. Collier, C. Nobata, and J. I. Tsuji. 2000. Building an annotated corpus from biology research papers. In *Proc. COLING 2000 Workshop on Semantic Annotation and Intelligent Content*, Luxembourg.

Tateisi, Y. and J. I. Tsuji. 2004. Part-of-speech annotation of biological abstracts. In *Proc. 4th LREC*, Barcelona.

Teufel, S., J. Carletta, and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proc. 8th EACL*, pages 110–117.

Teufel, S. and M. Moens. 2002. Summarising scientific articles–experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

Traum, D. R. and E. A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3). Special Issue on Non-literal Language.

Uebersax, John S. 1988. Validity inferences from interobserver agreement. *Psychological Bulletin*, 104(3):405–416.

Uebersax, John S. and William M. Grove. 1990. Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9:559–572.

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

Véronis, J. 1998. A study of polysemy judgments and inter-annotator agreement. In *Proc. of SENSEVAL-1*.

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, Maryland.

Vlachos, A., N. Karamanis, R. Seal, I. Lewin, C. Yamada, C. Gasperin, and T. Briscoe. 2006. *Annotation Guidelines for Named Entity Recognition in the FlySLIP Project*. University of Cambridge, CRL, Cambridge. Available from ....

Voorhees, E. and D. Harman. 1998. Overview of the seventh text retrieval conference (TREC-7). NIST special publication.

Wayne, C. 2000. Multilingual topic tracking and detection: successful research enabled by corpora and evaluation. In *Proc. 2nd LREC*, pages 1487–1493, Athens.

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