Evaluating discourse annotation: Some recent insights and new approaches

Jet Hoek
Utrecht Institute of Linguistics OTS
Utrecht University
j.hoek@uu.nl

Merel C. J. Scholman
Language Science and Technology
Saarland University
m.c.j.scholman
@coli.uni-saarland.de

Abstract

Annotated data is an important resource for the linguistics community, which is why researchers need to be sure that such data are reliable. However, arriving at sufficiently reliable annotations appears to be an issue within the field of discourse, possibly due to the fact that coherence is a mental phenomenon rather than a textual one. In this paper, we discuss recent insights and developments regarding annotation and reliability evaluation that are relevant to the field of discourse. We focus on characteristics of coherence that impact reliability scores and look at how different measures are affected by this. We discuss benefits and disadvantages of these measures, and propose that discourse annotation results be accompanied by a detailed report of the annotation process and data, as well as a careful consideration of the reliability measure that is applied.

1 Introduction

Linguistics researchers often make use of large amounts of data that are annotated by two or more coders. In order to draw conclusions from these data, researchers need to be sure that such data are reliable. Reliability "is the extent to which different methods, research results, or people arrive at the same interpretations or facts" (Krippendorff, 2011); data are reliable if coders agree on the labels assigned to, for instance, discourse relations (Artstein and Poesio, 2008). One way in which the reliability of annotated data can be measured is by calculating the inter-coder agreement: a numerical index of the extent of agreement between the coders.

Spooren and Degand (2010) note that sufficiently reliable annotation appears to be an issue within the field of discourse coherence. As the main reason for this, they point to the fact that coherence is a feature of the mental representation that readers form of a text, rather than of the linguistic material itself. Discourse annotation thus relies on coders' interpretation of a text, which makes it a particularly difficult task. This idea is for instance supported by studies that show that coders tend to agree more when annotating explicit coherence relations, which are signalled by a connective or cue phrase (*because, for this reason*), than when annotating implicit coherence relations, which contain no or less linguistic markers on which coders can base their decision (e.g., Miltsakaki et al., 2004; Prasad et al., 2008). Spooren and Degand (2010) argue that low agreement scores may contribute to the fact that reliability scores are often not reported in corpus-based discourse studies. They discuss several possible solutions to increase the reliability of discourse annotation tasks, including providing the annotators with more training, improving annotation protocols, and changing the definition of what a good or sufficient agreement score is.

Since Spooren and Degand (2010), there have been several new developments both in the discussion on inter-coder agreement measurement and within the field of discourse. In this paper, we address some of these insights.[1] First, we discuss a relatively new agreement measure, AC_1 (Gwet, 2002), that has

[1]It should be noted that although the focus of this paper will be on discourse-annotated data, some of the data characteristics we discuss are by no means unique to discourse, and all measures discussed in this paper could be used to calculate agreement for annotated data from different types of linguistic research as well.

been gaining popularity in recent years, and explore its suitability for measuring inter-coder agreement within the field of discourse annotation. AC_1 was introduced to solve some of the problems that Cohen's Kappa, the inter-coder agreement measure that is most widely used, presents.[2] Specifically, Kappa's values are sometimes relatively low, despite a high percentage of observed agreement; a problem known as the "Kappa paradox" (Feinstein and Cicchetti, 1990). As we will elaborate on in the next sections, this paradox occurs because Kappa is sensitive to certain characteristics of data that are very typical of discourse data.

After discussing AC_1 as a potential alternative for Cohen's Kappa in measuring the agreement between two (or more) expert coders, we briefly discuss some of the new methods of annotating discourse that have recently been used. These methods all involve the use of multiple naive, non-expert coders. Using non-expert coders is an attractive alternative to the conventional two-or-more expert coder scenario, since it allows researchers to obtain a lot of annotated data without extensive training sessions in a relatively fast and cheap way, especially when making use of crowdsourcing. For such annotation approaches, other methods for evaluating the reliability and quality of the annotations have been proposed.

2 Inter-coder agreement in (discourse) annotation

The discourse community makes frequent use of manually-annotated data, making inter-coder reliability a highly relevant issue for this field. However, a lot of research into reliability has been conducted by researchers from other fields, such as the medical field. These hypotheses and statistical measures have then been applied to discourse data, but differences between fields might affect the interpretation of agreement scores, as well as the appropriateness of a measure. For example, to interpret Kappa, researchers from all fields make use of Landis and Koch (1977)'s scale, which was originally designed for the medical field. Hripcsak and Heitjan (2002, p.101), however, argue that intermediate levels of Kappa cannot be interpreted consistently between fields or even within fields, because the interpretation relies heavily on the type of task and categories, the purpose of the measurement, and the definition of chance. In this section, we discuss specific characteristics of tasks and categories in the discourse coherence field, but first we address what sets apart linguistic annotation from other types of annotation, in order to highlight why different assumptions regarding reliability might be appropriate depending on the field.

Linguistic annotation differs from annotation tasks in other fields such as medicine for several reasons. In the medical field, patients are diagnosed as positive or negative, i.e., often the only two categories are 'yes' and 'no.' A data point often has an ultimate truth (the patient has the disease or does not have the disease), which can often be determined via different 'diagnostics' and for which additional evidence can emerge over time (due to the developmental course of diseases, for example). In linguistics, however, annotation tasks often consist of multiple categories. A data point never has an ultimate truth; rather, in many tasks, linguistics researchers study gradient phenomena where there are no right answers (Munro et al., 2010) and where it is not uncommon for data to be ambiguous (a coherence relation can for instance be causal and temporal at the same time). Finally, disagreements seem to be more equal in linguistics than in medicine. In the medical field, a false negative is worse than a false positive, since diagnosing a sick patient as healthy is worse than diagnosing a healthy patient as sick (e.g., Cicchetti et al., 2017). In linguistics, however, one mistake is not worse than another. These differences between domains do not at all imply that annotation tasks in discourse are easier or more difficult than those in the medical field, but they can play a role in whether a specific measure is suitable for determining agreement between coders.

In the next sections, we look at specific characteristics of typical discourse annotation tasks that influence the result of agreement measures, namely the number of categories and the distribution of categories. We illustrate our arguments using examples from discourse coherence data. However, the

[2]Like Kappa, AC_1 requires a simple categorical rating system. Gwet (2002) proposed a second statistic, called AC_2, for ordered categorical rating systems. This measure can be used as an alternative to weighted Kappa.

same arguments are often valid for other types of discourse annotation, including coreference coding (e.g., Van Deemter and Kibble, 2000), Translation Spotting (e.g., Cartoni et al., 2013), semantic role labeling (e.g., Palmer et al., 2005) or determining a discourse relation's segment-specific properties (e.g., Andersson and Spenader, 2014; Li, 2014; Sanders et al., 2012). Determining agreement is also relevant for experimentally obtained data, as in for instance continuation tasks or paraphrase tasks. The uneven occurrence of categories is an issue relevant to all these tasks, while the varying number of categories used in annotation is relevant mostly to the annotation of coherence relations, both in natural language and experimental data.

2.1 Number of categories

When annotating discourse relations, coders use labels to represent the way in which text segments relate to each other. Several different discourse annotation frameworks have been proposed, all of which have a different relation inventory. Frameworks differ not only in the exact labels they use, but also in the *number* of relational categories they distinguish. The DISCOR corpus (Reese et al., 2007), annotated within the framework of Segmented Discourse Representation Theory (SDRT), for example, uses 14 relation labels, while the RST Discourse Treebank (Carlson et al., 2003) uses 72 relation labels. The large variability in the number of categories between frameworks can contribute to low comparability of reliability scores between annotation efforts. A larger number of labels can for instance lead to more rare categories, which can in turn result in a lower reliability score, as we will see in the next sections.

The number of subtypes distinguished within classes in a single framework may also differ. The Penn Discourse Treebank 2.0 (PDTB 2.0, Prasad et al., 2008), for example, has 42 distinct labels, ordered in a hierarchy of four classes with three levels. The framework distinguishes 3 third-level labels within the class of TEMPORAL relations, but 11 third-level labels within CONTINGENCY relations. Such differences can make reliability scores difficult to compare between relation types even in a single framework.

2.2 Uneven distribution of categories

Regardless of the number of relation labels used, an uneven distribution of categories seems to be a common characteristic of discourse annotation. Since discourse annotation generally uses natural language as its basis, the frequency of a specific label is influenced by the frequency of the type of relation it refers to. The distribution of categories in discourse annotation can be skewed in multiple ways. For example, causal relations occur more often in natural text than non-causal relations such as LIST (e.g., Prasad et al. 2007). In addition, texts are characterized by an uneven distribution of connectives, with some connectives being very frequent (e.g., *because*), and other occurring less often (e.g., *consequently*). Finally, the distribution of relation types that specific connectives mark can also vary. Relations signaled by *so* are for instance more often RESULT than PURPOSE (e.g., Andersson and Spenader, 2014). Uneven prevalence of categories also extends beyond coherence relations. When it comes to coreference patterns, for instance, pronouns more often refer to the subject than to the object of the previous sentence.

The distribution of categories is also not stable between different types of discourse. The prevalence of relation types has been shown to differ between language modes (e.g., between spoken and written discourse, Sanders and Spooren, 2015), text genres (e.g., Demirşahin et al., 2012), and connectives (e.g., Andersson and Spenader, 2014), and between implicit and explicitly marked relations (e.g., Asr and Demberg, 2012). Similarly, coreference patterns can vary depending on the context; in the presence of an Implicit Causality verb, upcoming pronouns may more often refer back to the object than to the subject of the sentence (e.g., Garvey and Caramazza, 1974). Such variability in category distribution can reduce the comparability of reliability measures between annotation efforts, when using the same framework or the same labels.

The differences between discourse annotation efforts in the number of categories that are distinguished and the uneven distribution of categories can influence a reliability statistic such as Kappa, as will be explained in the next section. The variability in the prevalence of categories makes measuring the

reliability of discourse annotations even more problematic, since it has a varying effect on the reliability scores of annotation efforts that have been done using the same relation inventory. Specifically, if the observed prevalence of items in one of the categories is low, then there is insufficient information in the data to judge coders' ability to discriminate items, and Kappa may underestimate the true agreement (Hripcsak and Heitjan, 2002). This then also complicates a comparison of reliability scores between discourse annotation frameworks, since it prevents us from determining something along the lines of a 'framework correction.' For these reasons, it is important that researchers in the field of discourse annotation understand the distribution of relations in their frameworks and the prevalence in their data, and know how the agreement measures that they apply to their annotations are affected by this data. Without such knowledge and the appropriate reporting of these qualities, agreement scores on different discourse annotation tasks cannot be compared and the reliability of the data cannot be evaluated properly.

3 Inter-coder agreement measures: Kappa and AC_1

The simplest measure of agreement between coders is the percentage of agreement, also known as the observed agreement. This measure, however, is often not suitable for calculating reliability, as it does not take into account chance agreement (Scott, 1959). Chance agreement occurs when one or both coders rate an item randomly. This type of agreement can inflate the overall agreement and should therefore not contribute to a measure of inter-coder reliability (Artstein and Poesio, 2008).

In order to get a reliable index of the extent of agreement between coders, observed agreement has to be adjusted for chance agreement. Since it cannot be known which agreements between coders occurred by chance and which agreements are real, the proportion of chance agreement must be estimated (Gwet, 2001). Kappa and AC_1 correct for chance agreement on the basis of the same idea, namely that the ratio between the observed agreement and the expected agreement reflects how much agreement beyond chance was in fact observed. This idea is expressed in the following formula, which in both cases results in a score between -1 and 1:

$$\kappa, AC_1 = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the observed agreement, or percentage agreement, and P_e is the agreement that would be expected if the coders were acting only by chance. The crucial difference between Kappa and AC_1 lies in the way in which they estimate the expected agreement (P_e), as they have different assumptions about the coding distributions. In this section, we introduce each measure in turn, highlighting the differences between the measures as well as the respective drawbacks. We then illustrate the difference between Kappa and AC_1's scores using different annotation scenarios. The example data in this section will be used to illustrate the agreement measures, and will be reported in a two-way contingency table such as Table 1. This table represents a two-coder reliability study involving coders A and B and two categories.

It should be noted that while Kappa and AC_1 both range between 1 (complete agreement) to -1 (complete disagreement), neither score comes with a fixed value at which agreement can be considered satisfactory; guidelines and conventions on the interpretation of these measures are formed over time and can differ between fields. As mentioned above, Kappa is often interpreted using the scale proposed by Landis and Koch (1977), in which for instance $0.41-0.6 =$ moderate agreement, $0.61-0.8 =$ substantial agreement, and $0.81-1 =$ almost perfect agreement, but the cut-off point for acceptable agreement in computational linguistics is commonly set at $\kappa = 0.67$ (Di Eugenio and Glass, 2004), whereas Artstein and Poesio (2008) recommend considering $\kappa > 0.8$ as an indication of sufficient annotation quality. While these guidelines are helpful, they have no theoretical basis (Ludbrook, 2002; Xie, 2013) and are themselves subject to evaluation.

3.1 Cohen's Kappa

Cohen's Kappa assumes that "random assignment of categories to items is governed by prior distributions that are unique to each coder and that reflect individual annotator bias" (Artstein and Poesio, 2008, p.

561). In Kappa, chance agreement, or the amount of agreement that would be expected if annotators were acting only by chance ("expected agreement"), is estimated using the marginal distribution (i.e., the probability that a category is used by the coders):

$$P_e(\kappa) = (f_1 \cdot g_1 + f_2 \cdot g_2)/N^2 \tag{2}$$

where f_1, f_2, g_1 and g_2 correspond to the marginal totals in Table 1.

	Coder A		
Coder B	1	2	Total
1	a	b	g1
2	c	d	g2
Total	f1	f2	*N*

Table 1: Coders and response categories

3.2 Gwet's AC_1

AC_1's definition of chance agreement is based on the premises that chance agreement occurs when at least one coder guesses and that only an unknown proportion of ratings is random. AC_1 thus assumes that coders' agreements are at least in part not due to chance. In addition, AC_1 takes into account the prevalence of the categories for its estimation of chance agreement (Gwet, 2001).

The calculation of chance agreement in AC_1 is expressed by the following formula:

$$P_e(AC_1) = \frac{1}{(K-1)} \sum_{q=1}^{K} \left(\frac{N_q}{N} \cdot \frac{N - N_q}{N} \right) \tag{3}$$

whereby K refers to the total number of categories and q to a specific category. N_q refers to the average number of times a certain category is used by a coder and is, in case of two coders, equivalent to $(f_q + g_q)/2$. N_q/N thus represents the percentage of items labeled as category q and $(N - N_q)/N$ represents the percentage of items not labeled as category q (see also Zhao et al., 2013). Hence, whereas Kappa's formula of chance agreement is based on the chance that Coder A and B both categorize an item as '1,' the chance that both coders categorize an item as '2', etc., AC_1's chance agreement formula is based on the chance that a certain category is used.

The values from AC_1's formula for chance agreement are crucially different from those of Kappa's chance agreement formula because AC_1 does not assume a prior individual coder bias. Instead, it is based on the possibility that one or both of the coders perform a random classification. As such, Gwet (2002, p. 3) argues that a reasonable value for chance agreement probability should not exceed 0.5. Consequently, AC_1's chance agreement caps the probability within 0–0.5, whereas Kappa's chance agreement probability can be anywhere between 0 and 1. AC_1's limit of 0.5 aims to prevent the occurrence of a similar erratic behaviour that leads to Kappa's paradoxes. In addition, AC_1's chance agreement, unlike Kappa's chance agreement, is positively correlated with the difficulty of a task, since it includes the chance of coders annotating randomly (Feng, 2015); as a task gets more difficult, chances that coders guess increase. The next section will explore Kappa's and AC_1's behavior in different annotation scenarios.

3.3 Kappa vs. AC_1 in annotation scenarios

Because Kappa bases its chance agreement on individual coder biases, the resulting agreement score can be greatly affected by the distribution of the categories in the data. Specifically, when the marginal distributions are imbalanced, the resulting κ is lower than when the marginal distributions are balanced. In Table 2, the distribution is symmetrical and balanced; the coders agree on an equal amount of items for both categories, and category 1 is used approximately as often as category 2. The observed agreement is 100/120 = 0.83, and the κ score for these data is 0.67 (see Table 6 for an overview of all scores).[3] In this

[3] All agreement scores reported in this paper were calculated using the R package agree.coeff2.r.

	Coder A		
Coder B	1	2	Total
1	50	12	62
2	8	50	58
Total	58	62	120

Table 2: Symmetrical balanced distribution

	Coder A		
Coder B	1	2	Total
1	20	12	32
2	8	80	88
Total	28	92	120

Table 3: Symmetrical imbalanced distribution

	Coder A		
Coder B	1	2	Total
1	0	12	12
2	8	100	108
Total	8	112	120

Table 4: Symmetrical imbalanced distribution with empty target cell

	Coder A		
Coder B	1	2	Total
1	5	110	115
2	0	5	5
Total	5	115	120

Table 5: Highly imbalanced distribution in opposite direction

scenario, AC_1's chance agreement is the same. Consequently, the agreement scores are also the same. Hence, when the data is distributed evenly, Kappa and AC_1 give comparable scores.

In Table 3, the distribution of categories is also approximately the same for each coder (i.e., 30/90), but category 2 is used more often than category 1, and the distribution is therefore symmetrical but imbalanced. For these types of distributions, AC_1 and Kappa yield different scores. Kappa assumes that both coders have a bias toward category 2 and that, as such, they would agree often if they guessed according to their biases. The observed agreement is the same as for the data in Table 2 (0.83), but Kappa's chance agreement is higher (0.62), resulting in a lower κ score (0.56).[4] AC_1, by contrast, assumes that uneven categories are a property of the data. Consequently, AC_1 assumes a lower value for chance agreement than Kappa (0.38), which results in a higher agreement score (0.73).

Table 4 illustrates a scenario in which two coders have reached a high observed agreement (0.83), but have not managed to agree on a single case for category 1 (resulting in an empty target cell). The κ score for this annotation task is -0.09, which indicates that agreement was around chance level. Kappa estimates chance agreement at 0.85. This is an extreme case of a low Kappa score for a task with a high observed agreement. The fact that the Kappa score is much lower than the observed agreement is in this case not completely unreasonable, since even though both coders used category 1 several times, they did not agree on a single case for this category. On the other hand, having a reliability score around chance implies that coders did no better than if they were guessing, even though it seems plausible to assume that at least part of the items that were classified as category 2 were assigned this label because the coders were certain that the item belonged to this category.

Because AC_1, unlike Kappa, takes into account both the number of categories and the prevalence of those categories, the chance agreement is much lower than Kappa's (0.22) and the resulting reliability score is higher (0.80). Note that AC_1's estimation of chance agreement for Table 4 is lower than for Tables 2 and 3, and that its reliability score for 4 is therefore higher than for the other two scenarios. This can be considered counter-intuitive; after all, there is only one category in Table 4 on which the coders have managed to agree. The coders have not been able to reliably assign any items to the other category, which constitutes 50% of the categories in a 2×2 table. One would expect that the corresponding agreement score is affected by this. Zhao et al. (2013) note that this is an abnormality in AC_1: in case of a very skewed distribution with an empty target cell, AC_1 turns out higher than what seems justified.

A similar abnormality in AC_1 is that unused categories influence the reliability score. For instance,

[4]Sometimes, KappaMAX is used to correct Kappa in case of uneven categories. KappaMAX is calculated using the same formula as Kappa, but the '1' in formula 1 is replaced by the maximum value for observed agreement possible (if f_1 is the smallest marginal total, $maxp_o = (f_1 + g_2)/N$; if f_2 is the smallest marginal total, $maxp_o = (f_2 + g_1)/N$ (see also Feinstein and Cicchetti, 1990)). Although KappaMAX can correct Kappa's prevalence problem, it has been reported to overcorrect in case of coder bias (Feinstein and Cicchetti, 1990).

Table	P_o	Kappa		AC_1	
		P_e	κ	P_e	AC_1
2	0.83	0.50	0.67	0.50	0.67
3	0.83	0.62	0.56	0.38	0.73
4	0.83	0.85	-0.09	0.22	0.80
5	0.08	0.08	0.004	0.50	-0.83

Table 6: Values for observed agreement, chance agreement and reliability scores for Tables 2-5.

if we were to add an empty category ('3') to Table 2, the AC_1 score would rise from 0.67 to 0.78. Kappa, by contrast, is not affected by the unused category, and gives a score of 0.67 in both cases. This may be perceived as a positive feature of AC_1, since the measure can reflect that coders have successfully not attributed any of the items to a certain category. On the other hand, it makes AC_1 vulnerable to inflation through the inclusion of useless categories in an annotation task. It should, however, be noted that the inflation effect of unused categories decreases as the number of used categories increases.

Instead of being low relative to the observed agreement, Kappa can also be high. Table 5 presents an extreme case of coder disagreement. The observed agreement is very low (0.08), but κ is 0.0004, which suggests that agreement was around chance. Looking at the table, however, it appears that there is almost perfect disagreement. It could be argued that this too is a type of agreement; even though the coders did not use the same label, they did make the same categorization of the data. The agreement should therefore close to -1. AC_1's agreement score for the data in Table 5 (0.80) therefore much better reflects the almost perfect disagreement that the coders showed. Although such extreme cases of disagreement are rare in annotation, this example demonstrates Kappa's potential to be relatively high when coders disagree on many items. Ideally, a measure would be able to deal properly with all possible scenarios, including one of almost perfect disagreement.

3.4 Using AC_1 to evaluate discourse annotations

As discussed in Section 2, skewed data are fairly common in discourse annotation tasks and distributions can vary depending on the context or the task. This variation complicates a comparison of reliability scores between annotation efforts. In addition, discourse frameworks often have many categories and the prevalence of these categories in a text or dataset is unknown. Empty categories are therefore very likely to occur in discourse annotation tasks. Disagreements on a rare category can have a big impact on the Kappa score, especially when the target cell for that category remains empty. AC_1 is more robust to skewedness and variability in the distribution of categories, and therefore seems promising as a measure for evaluating agreement in discourse annotation. Results from several studies and simulations have suggested that AC_1 is a reliable alternative measure for calculating inter-coder agreement (e.g., Gwet, 2001; Wongpakaran et al., 2013; Xie, 2013). Moreover, AC_1 has been applied often in the medical field (e.g., Bryant et al., 2013; Crowle et al., 2017; Fuller et al., 2017; Marks et al., 2016) and has also been used in the computational linguistics field (Besser and Alexandersson, 2007; Haley, 2009; Hillard et al., 2007; Kranstedt et al., 2006; Purpura and Hillard, 2006; Yang et al., 2006), but no research in the field of discourse annotation has used AC_1 as of yet.

It is important that researchers are aware that both Kappa and AC_1 behave abnormally under some conditions. Zhao et al. (2013) point out that we cannot be entirely sure exactly when a measure like AC_1 – which assumes that coding happens randomly only part of the time – overestimates reliability or by how much and, vice versa, when a measure like Kappa – which assumes maximum-randomness – underestimates reliability. The choice for any agreement statistic should be well-motivated and researchers should be transparent about the distributions in their data. It might also be warranted that the guidelines for what constitutes satisfactory agreement are slightly stricter for AC_1 compared to those for Kappa, whether they be 'formalized' guidelines such as Landis and Koch (1977), framework-specific guidelines, or practices developed over annotation efforts.

Since AC_1 is still a relatively new agreement measure, it is possible that more frequent use and more examination will uncover more issues. We encourage discourse researchers to consider using both AC_1 and Kappa, and to be explicit about the characteristics of their data that might influence the suitability of their inter-coder agreement measure. Regardless of which measure researchers choose for their data, we advise them to include contingency tables to make annotation results more transparent and to allow readers to evaluate the results as well.

4 Multiple coders and crowdsourcing

Traditional annotation tasks consist of two expert coders. However, as Krippendorff (2004) notes, using more, non-expert coders can help ensure the reliability of the annotated data. In recent years, studies have begun to explore whether non-expert, non-trained (also referred to as naive) coders can also be employed for discourse annotation tasks (compared to expert coders). There are several advantages in employing such coders: non-experts are easier to come by, making it easier to employ a large number. Multiple annotators reduce the risk of coder bias in the data (Artstein and Poesio, 2005). Moreover, employing non-expert coders allows for a cost-effective and fast approach to collecting large amounts of data.

For non-expert annotations to be valuable, researchers have to be sure that they are sufficiently reliable (compared to expert annotations). There are several ways to evaluate annotations generated by non-expert, non-trained coders. For example, coders can be compared to each other based on their performance (Peldszus and Stede, 2013). Alternatively, they can be compared to a gold standard developed by an expert (Scholman et al., 2016). Typically, an adapted version of Kappa (i.e., Fleiss' Kappa, Davies and Fleiss, 1982) is used to calculate agreement for tasks with multiple coders, but AC1 could in fact also be used. Recall, precision, and F-scores can also provide valuable insights into problematic categories in the framework that is used.

To facilitate crowdsourced annotation projects without a gold standard set by experts, new methods of coding evaluation have been proposed, such as models that can extract a gold standard from crowdsourced data. Aroyo and Welty (2013), for instance, propose creating binary annotation vectors for all annotated items. These vectors then function as a gold standard to which individual annotations can be compared: comparing individual coder vectors to the total item vectors (minus the data supplied by that coder) gives an indication of coder disagreement, or the quality of each individual coder, whereas comparing all coder vectors for a single item to the averaged item vector functions as a measure of sentence clarity, or sentence ambiguity (for details, see Aroyo and Welty, 2013).

Another, more commonly used method is an approach using probabilistic item-response models that draw inferences about annotated data (Hovy et al., 2013; Passonneau and Carpenter, 2014). Such models use unsupervised learning to estimate the probability of labels for every item and coder. The utility of such a model lies in its ability to support meaningful inferences from the data, such as an estimate of the true prevalence of each category. Specifically, two features of probabilistic models make them an attractive alternative to more traditional reliability measurement methods. First, the models allow researchers to differentiate between coders; specifically, they can adjust for annotations from noisy coders, since some coders perform better than others. This is for instance done by giving different weights to annotators that answer correctly less often than others (Hovy et al., 2013). Second, probabilistic models cannot only identify the correct label for an item based on the crowdsourced annotations, they can also provide a confidence measure that indicates how likely it is that this label is indeed the correct label (cf. Hovy et al., 2013; Passonneau and Carpenter, 2014). This allows researchers to balance between coverage, i.e., the amount of data that is annotated, with accuracy, i.e., the trustworthiness of each annotation; as Hovy et al. 2013 explain, researchers can favor a different trade-off between coverage and accuracy depending on their research purposes. With the exception of Kawahara et al. (2014), no work has evaluated crowdsourced discourse relation annotations using probabilistic models. This seems a promising topic for future research.

The use of multiple (naive) coders also opens up other possibilities for representing the data. It allows researchers to study the distribution of responses over many coders, rather than specific data points

(Munro et al., 2010). This can be beneficial in unsupervised approaches where it is assumed that there is no one ground truth. Rohde et al. (2016) and Scholman and Demberg (2017), for example, present confusion matrices, percentage agreement and distribution plots to show that, often, multiple interpretations are possible for a single discourse relation. Rohde et al. (2016) argue that without gathering judgements from a crowd of coders, differences in annotation might be written off as coder error or bias, or a low level of inter-coder agreement. Based on their crowdsourced data, they conclude that disagreements on the interpretation of certain relations might be due to the fact that not every item can be assigned one right answer. Using the distribution of responses from multiple coders to determine whether disagreements are due to biases or errors, or caused by genuine ambiguity or double meanings could in the future lead to valuable insights for the evaluation of discourse annotation efforts by a limited number of expert annotators as well, especially if we can determine a subset of relations (or relation characteristics) that tend to allow multiple interpretations.

While there are many benefits to crowdsourcing annotations, using a large number of naive coders to annotate discourse relations may not be without its difficulties. As discussed in Sections 1 and 2, annotating discourse relations is a highly complicated task. Expert annotators usually spend a long time developing or acquainting themselves with an annotation framework and its relation inventory, annotation manuals tend to be very extensive, and annotation tasks often involve practice phases and discussion. Replicating this process in a crowdsourcing setting may be difficult, if not inconceivable. Instead of trying to use existing annotation manuals and procedures, however, researchers should consider developing methods that allow them to reap the benefits of crowdsourcing, while at the same time approximating the results yielded by a traditional annotation scenario. They may, for instance, opt for connective/cue phrase insertion tasks (cf. Rohde et al., 2016; Scholman and Demberg, 2017), in which case the connectives coders can choose from should be reliably associated with a specific type of relation. In addition, the annotation process could be simplified by cutting it up into several different steps, as in Scholman et al. (2016), or by including only a limited set of relations, as in Kawahara et al. (2014). Alternative solutions could be training coders to annotate only a small subset of relations, such as temporal relations, or teaching them to annotate only a single distinction, for instance the difference between RESULT and PURPOSE relations, between contrastive and temporal *while*, or between inclusive and exclusive DISJUNCTION relations.

5 Conclusion

This paper reviewed some recent developments concerning reliability evaluation within linguistic annotation in general and discourse annotation in specific. We explored the suitability of a relatively new agreement measure, AC_1, to evaluate the reliability of discourse annotation. This measure could be considered as a possible alternative for, or be used in addition to Cohen's Kappa. In general, the comparison demonstrated how agreement statistics can be influenced by properties of the data. We also discussed some annotation methods that have been used as alternatives to the two-or-more expert coders procedure, how the reliability can be determined for these methods, and how findings from these studies could help further our understanding of the practice of discourse annotation.

When reporting the results of a study that involves annotation, it is advisable to be transparent about the annotation process and to carefully consider which agreement measure is reported. Reporting (multiple) agreement scores and making raw annotation data available would facilitate other researchers to judge the reliability of the annotated data and, consequently, the findings of a study. In addition, it would enable a comparison between different annotation efforts and frameworks.

Acknowledgements

This research was funded by the SNSF Sinergia project MODERN (CRSII2_147653) and the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding". We are grateful to the anonymous reviewers for their helpful suggestions.

References

Andersson, M. and J. Spenader (2014). Result and purpose relations with and without 'so'. *Lingua 148*, 1–27.

Aroyo, L. and C. Welty (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM 2013*.

Artstein, R. and M. Poesio (2005). Bias decreases in proportion to the number of annotators. *Proceedings of the Conference on Formal Grammar and Mathematics of Language (FG-MoL)*, 141–150.

Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34*(4), 555–596.

Asr, F. T. and V. Demberg (2012). Implicitness of discourse relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 2669–2684. Citeseer.

Besser, J. and J. Alexandersson (2007). A comprehensive disfluency model for multi-party interaction. In *Proceedings of SigDial*, Volume 8, pp. 182–189.

Bryant, J., L. E. Skolarus, B. Smith, E. E. Adelman, and W. J. Meurer (2013). The accuracy of surrogate decision makers: Informed consent in hypothetical acute stroke scenarios. *BMC Emergency Medicine 13*(1), 18.

Carlson, L., D. Marcu, and M. E. Okurowski (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, pp. 85–112. Springer.

Cartoni, B., S. Zufferey, and T. Meyer (2013). Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue and Discourse 4*(2), 65–86.

Cicchetti, D. V., A. Klin, and F. R. Volkmar (2017). Assessing binary diagnoses of bio-behavioral disorders: The clinical relevance of Cohen's Kappa. *The Journal of Nervous and Mental disease 205*(1), 58–65.

Crowle, C., C. Galea, C. Morgan, I. Novak, K. Walker, and N. Badawi (2017). Inter-observer agreement of the general movements assessment with infants following surgery. *Early Human Development 104*, 17–21.

Davies, M. and J. L. Fleiss (1982). Measuring agreement for multinomial data. *Biometrics 38*(4), 1047–1051.

Demirşahin, I., A. Sevdik-Çallı, H. Ö. Balaban, R. Çakıcı, and D. Zeyrek (2012). Turkish discourse bank: Ongoing developments. In *Proceedings of the First Workshop on Language Resources and Technologies for Turkish Languages*, pp. 15–19. Citeseer.

Di Eugenio, B. and M. Glass (2004). The kappa statistic: A second look. *Computational linguistics 30*(1), 95–101.

Feinstein, A. R. and D. V. Cicchetti (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology 43*(6), 543–549.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology 11*(1), 14–22.

Fuller, G., S. Kemp, and M. Raftery (2017). The accuracy and reproducibility of video assessment in the pitch-side management of concussion in elite rugby. *Journal of science and medicine in sport 20*(3), 246–249.

Garvey, C. and A. Caramazza (1974). Implicit causality in verbs. *Linguistic Inquiry 5*(3), 459–464.

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS Publishing Company.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment 1*(6), 1–6.

Haley, D. (2009). *Applying latent semantic analysis to computer assisted assessment in the computer science domain: A framework, a tool, and an evaluation*. Ph. D. thesis, The Open University.

Hillard, D., S. Purpura, and J. Wilkerson (2007). An active learning framework for classifying political text. In *Annual Meeting of the Midwest Political Science Association, Chicago*.

Hovy, D., T. Berg-Kirkpatrick, A. Vaswani, and E. H. Hovy (2013). Learning whom to trust with mace. In *HLT-NAACL*, pp. 1120–1130.

Hripcsak, G. and D. F. Heitjan (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics 35*(2), 99–110.

Kawahara, D., Y. Machida, T. Shibata, S. Kurohashi, H. Kobayashi, and M. Sassano (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 269–278.

Kranstedt, A., A. Lücking, T. Pfeiffer, H. Rieser, and M. Staudacher (2006). Measuring and reconstructing pointing in visual contexts. In *Proceedings of the Brandial*, pp. 82–89.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research 30*(3), 411–433.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures 5*(2), 93–112.

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

Li, F. (2014). *Subjectivity in Mandarin Chinese: The meaning and use of causal connectives in written discourse*. Netherlands Graduate School of Linguistics. Ph.D. Dissertation.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology 29*(7), 527–536.

Marks, D., T. Comans, M. Thomas, S. K. Ng, S. O'Leary, P. G. Conaghan, P. A. Scuffham, and L. Bisset (2016). Agreement between a physiotherapist and an orthopaedic surgeon regarding management and prescription of corticosteroid injection for patients with shoulder pain. *Manual Therapy 26*, 216–222.

Miltsakaki, E., R. Prasad, A. K. Joshi, and B. Webber (2004). The Penn Discourse TreeBank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.

Munro, R., S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–130. Association for Computational Linguistics.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics 31*(1), 71–106.

Passonneau, R. J. and B. Carpenter (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics 2*, 311–326.

Peldszus, A. and M. Stede (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 7*(1), 1–31.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. Webber (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, A. K. Joshi, L. Robaldo, and B. Webber (2007). *The Penn Discourse Treebank 2.0 annotation manual*. University of Pennsylvania.

Purpura, S. and D. Hillard (2006). Automated classification of congressional legislation. In *Proceedings of the 2006 International Conference on Digital government Research*, pp. 219–225. Digital Government Society of North America.

Reese, B., J. Hunter, N. Asher, P. Denis, and J. Baldridge (2007). *Reference manual for the analysis and annotation of rhetorical structure (version 1.0)*. Technical report. Austin: University of Texas, Departments of Linguistics and Philosophy. Available online: http://timeml. org/jamesp/annotation_manual. pdf.

Rohde, H., A. Dickinson, N. Schneider, C. N. Clark, A. Louis, and B. Webber (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*, pp. 49–58.

Sanders, T. J. and W. P. Spooren (2015). Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. *Linguistics 53*(1), 53–92.

Sanders, T. J., K. Vis, and D. Broeder (2012). Project notes of CLARIN project DiscAn: Towards a discourse annotation system for Dutch language corpora. In *Workshop on Interoperable Semantic Annotation (ISA)*, pp. 61–65.

Scholman, M. C. and V. Demberg (2017). Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse 8*(2), 56–84.

Scholman, M. C., J. Evers-Vermeul, and T. J. Sanders (2016). Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse 7*(2), 1–28.

Scott, W. A. (1959). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly 19*(3), 321–325.

Spooren, W. P. and L. Degand (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory 6*(2), 241–266.

Van Deemter, K. and R. Kibble (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics 26*(4), 629–637.

Wongpakaran, N., T. Wongpakaran, D. Wedding, and K. Gwet (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology 13*(61), 1–7.

Xie, Q. (2013). Agree or disagree? A demonstration of an alternative statistic to Cohen's kappa for measuring the extent and reliability of agreement between observers. Unpublished manuscript.

Yang, H., J. Callan, and S. Shulman (2006). Next steps in near-duplicate detection for eRulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*, pp. 239–248. Digital Government Society of North America.

Zhao, X., J. S. Liu, and K. Deng (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association 36*(1), 419–480.