Inter-annotator Agreement

Ron Artstein

Abstract

This chapter touches upon several issues in the calculation and assessment of inter-annotator agreement. It gives an introduction to the theory behind agreement coefficients and examples of their application to linguistic annotation tasks. Specific examples explore variation in annotator performance due to heterogeneous data, complex labels, item difficulty, and annotator differences, showing how global agreement coefficients may mask these sources of variation, and how detailed agreement studies can give insight into both the annotation process and the nature of the underlying data. The chapter also reviews recent work on using machine learning to exploit the variation among annotators and learn detailed models from which accurate labels can be inferred. I therefore advocate an approach where agreement studies are not used merely as a means to accept or reject a particular annotation scheme, but as a tool for exploring patterns in the data that are being annotated.

Keywords

Inter-annotator agreement · Kappa · Krippendorff's alpha · Annotation reliability

R. Artstein (✉)
Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Playa Vista, CA, USA
e-mail: artstein@ict.usc.edu

N. Ide and J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*,
DOI 10.1007/978-94-024-0881-2_11

1 Why Measure Inter-Annotator Agreement

It is common practice in an annotation effort to compare annotations of a single source (text, audio etc.) by multiple people. This is done for a variety of purposes, such as validating and improving annotation schemes and guidelines, identifying ambiguities or difficulties in the source, or assessing the range of valid interpretations (not to mention the study of annotation in its own right). The comparison may take a variety of forms, for instance a qualitative examination of the annotations, calculation of formal agreement measures, or statistical modeling of annotator differences. What is common to these various studies is the realization that there exists variation in annotator performance, and this variation needs to be examined in order to understand what the annotators are doing, and to be able to make meaningful use of the annotators' output. This chapter will concentrate on formal means of comparing annotator performance.

The textbook case for measuring inter-annotator agreement is to assess the reliability of an annotation process, as a prerequisite for ensuring correctness of the resulting annotations. The reasoning is as follows. The annotation scheme, as envisioned by the experimenter and codified in the annotation guidelines, defines (or is intended to define) a correct annotation for each particular source. Since the actual annotations are created by the annotators, there is no reference corpus against which the annotations can be checked for correctness. In lieu of correctness of the annotated corpus, then, we check for reliability of the annotation process, which serves as a necessary (but not sufficient) condition for correctness: if the annotation process is not reliable, then we cannot expect the annotations to be correct. An annotation process is reliable if it is reproducible, that is if the annotations yield consistent results. To check for consistency we need to apply the annotation process several times to the same source, and we need to use different annotators because a single person might remember their annotations from a previous round. Agreement among annotators on the same source data gives a measure of the extent to which the annotation process is consistent, or reproducible.

Rationale for measuring agreement
Agreement among annotators
↓ *demonstrates*
Reliable annotation process
↓ *necessary but not sufficient for*
Correct annotations

Reliability is typically assessed on a sample of the material to be annotated, the idea being that once the process is demonstrated to be reliable, it can be applied to the remainder of the material by just one annotator. Several conditions need to be met in order for agreement to be taken as an indication for reliability (see [24]). The annotators should follow written guidelines, to make sure that the annotation process relies on knowledge that is transferable. They must work independently, so

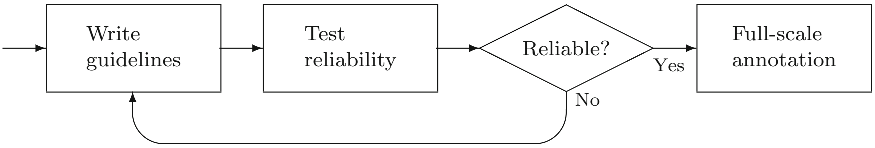

Fig. 1 Iterative reliability testing

that agreements come from a shared understanding of the annotation guidelines rather than individual discussions on case points. Annotators should be drawn from a well-defined population in order for the researchers to know what shared assumptions they bring to the annotation process prior to reading the guidelines. The sample material must be representative of the totality of the material in terms of the annotated phenomena. And not any measure of agreement will do: Sect. 2 will introduce the accepted ways of measuring agreement in a way that reflects reliability.

Agreement testing is part of an iterative methodology for developing reliable annotation schemes. The standard procedure is to develop a scheme, test it for reliability, analyze the test results to revise the scheme, and iterate until the desired level of reliability has been reached – at which point, full-scale annotation can proceed (Fig. 1). However, reliability is not uniform, and an annotation scheme that is reliable overall may be unreliable with respect to certain parts of the data or distinctions within the data. Section 3 illustrates some of the ways reliability can vary within an annotation effort; it also shows how agreement measures can be used for analysis beyond annotation scheme validation.

While reliable annotation is a desirable goal, it is often quite difficult to attain in linguistic annotation tasks. Less-than-reliable annotation may in some cases contain sufficient information to allow inference of the correct labels, by learning models for the annotators and the annotations they produce. Such applications still require data from multiple annotators in order to learn the models; these applications are explored in Sect. 4.

2 Standard Measures: The Kappa/Alpha Family

In a prototypical annotation task, annotators assign *labels* to specific *items* (words, segments etc.) in the source. The simplest way to measure agreement between annotators is to count the number of items for which they provide identical labels, and report that number as a percentage of the total to be annotated. This is called **raw agreement** or **observed agreement**, and according to Bayerl and Paul [5] it is still the most common way of reporting agreement in the literature. Raw agreement is easy to measure and understand; however, agreement in itself does not imply that the annotation process is reliable, because some agreement may be accidental – and this accidental agreement could be very high. This is especially clear when annotating for sparse phenomena, for example the task of identifying gene-renaming sequences in text, as presented in Fort et al. [15]: out of over 19,000 tokens in the source, only about 200 (1%) represent gene-renaming sequences. If two annotators each identified a completely different set of 200 tokens, they would still agree that 98% of the data do not represent gene-renaming sequences; but this agreement does not demonstrate that the annotation results are reproducible, or reliable.

The accepted way to measure meaningful agreement, which implies reliability, is by using a coefficient from the kappa/alpha family (I use this name because these are the most familiar coefficients of this type). These coefficients are intended to calculate the amount of agreement that was attained above the level expected by chance or arbitrary coding. Let $\mathrm{A_o}$ denote the amount of observed inter-annotator agreement (a number between 0 and 1), and let $\mathrm{A_e}$ be the level of agreement expected by some model of arbitrary coding (more on this later). The amount of agreement above chance is $\mathrm{A_o} - \mathrm{A_e}$ (this could be negative, if agreement is below chance expectation); the maximum possible agreement above chance is $1 - \mathrm{A_e}$. The ratio between these quantities is a coefficient whose value is 1 when agreement is perfect, and 0 when agreement is at chance level.

$$\kappa, \pi, \ldots = \frac{\mathrm{A_o} - \mathrm{A_e}}{1 - \mathrm{A_e}} \tag{1}$$

Many coefficients belong to the above paradigm; among the early proposals are S [6], π [31] and κ [9], which were followed by numerous extensions. I consider α [23] to be part of this family even though it has somewhat different roots and is expressed in terms of disagreement rather than agreement.

$$\alpha = 1 - \frac{\mathrm{D_o}}{\mathrm{D_e}} \tag{2}$$

Equations 1 and 2 are equivalent if disagreement is taken to be the complement of agreement, that is $\mathrm{D_o} = 1 - \mathrm{A_o}$ and $\mathrm{D_e} = 1 - \mathrm{A_e}$. The advantage of expressing the coefficient in terms of disagreement is that it allows expressing the extent of disagreement in units other than percentages, when such units make sense.

The main difference between the various coefficients is in how they conceptualize the notion of agreement expected by arbitrary coding, and therefore how they calculate the chance component of the equation. The debates on the matter have been raging for decades, in particular on how to treat individual differences between annotators (see for example [7,10,12,14,18,22,25,33]). A brief review of these issues is given in Artstein and Poesio [1, Sect. 3], and I do not see a need to revisit the matter here. I will therefore proceed with the coefficients that are the most appropriate for gauging the reliability of the annotation process, that is Fleiss's κ and Krippendorff's α.

Note: The term "kappa" (κ) may refer to several distinct agreement coefficients, most commonly those of Cohen [9] and Fleiss [13]. These coefficients are not compatible, as they use distinct conceptions of agreement expected by chance (Fleiss's κ is more closely related to Scott's π, and was referred to as multi-π in Artstein and Poesio [1]). When reporting a result using "κ" it is important to clarify which coefficient is being used.

Observed agreement in Fleiss's κ is defined in the spirit of the characterization given at the beginning of this section: the proportion of items on which two annotators agree. When there are more than two annotators, observed agreement is calculated pairwise. Let $\mathbf{c}$ be the number of annotators, and let $\mathbf{n}_{ik}$ be the number of annotators who annotated item i with label k. For each item i and label k there are $\binom{\mathbf{n}_{ik}}{2}$ pairs of annotators who agree that the item should be labeled with k; summing over all the labels, there are $\sum_k \binom{\mathbf{n}_{ik}}{2}$ pairs of annotators who agree on the label for item i, and agreement on item i is the number of agreeing pairs divided by the total number of pairs of annotators $\binom{\mathbf{c}}{2}$. Overall observed agreement is the mean agreement per item, that is the sum of observed agreement for each item i divided by the total number of items $\mathbf{i}$.

$$\text{[Fleiss's } \kappa\text{]} \qquad \mathrm{A_o} = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_i \sum_k \mathbf{n}_{ik}(\mathbf{n}_{ik}-1) \tag{3}$$

Krippendorff's α is similar to Fleiss's κ, but while κ treats all disagreements as equally severe, α incorporates a distance function that sets a specific level of disagreement for each pair of labels. For example, if the annotators' labels denote intervals on a numerical scale, as in magnitude estimation tasks, then the interval distance metric is appropriate, where for every pair of labels a and b, the distance $\mathbf{d}_{ab} = (a-b)^2$. Observed disagreement is calculated by counting the disagreeing pairs of judgments (rather than the agreeing pairs), and scaling each disagreement by the appropriate distance.

$$\text{[Krippendorff's } \alpha\text{]} \qquad \mathrm{D_o} = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_i \sum_{k_1} \sum_{k_2} \mathbf{n}_{ik_1}\mathbf{n}_{ik_2}\mathbf{d}_{k_1k_2} \tag{4}$$

When all labels are considered equally different from one another, the nominal distance metric is appropriate, where $\mathbf{d}_{ab} = 0$ if $a = b$, 1 if $a \neq b$. In this case, observed disagreement of Krippendorff's α (Eq. 4) is the exact complement of the observed agreement of Fleiss's κ (Eq. 3).

The chance component in a chance-corrected coefficient reflects the amount of agreement that would be attained if the annotators were making arbitrary annotations; such arbitrary annotations need not be uniform – as we saw in the gene-renaming annotation task above, an arbitrary annotation can be highly skewed and lead to high levels of chance agreement. In the absence of a priori knowledge of the annotators' propensity towards specific labels, we estimate this propensity from the annotated data. Since the reliability of an annotation procedure is independent of the actual annotators used, we abstract over individual annotator differences by using the totality of judgments from all annotators to calculate the distribution of labels. This is not to imply that any particular annotator works according to this distribution, or makes any arbitrary judgments; it is just used to calculate how much agreement we would expect to find *if* two arbitrary annotators were to make arbitrary annotations on the data.

Expected agreement according to Fleiss's κ is calculated as follows. Let $\mathbf{n}_k$ be the total number of labels of category k given by all the annotators, and let $N = \sum_k \mathbf{n}_k$ be the total number of labels given to the annotated data by all the annotators. The probability that an arbitrary annotator will make an arbitrary choice of category k is taken to be the proportion of k labels among all the labels, that is $\frac{1}{N}\mathbf{n}_k$. Therefore the probability that two arbitrary annotators, making arbitrary choices, will happen to agree on category k is taken to be $(\frac{1}{N}\mathbf{n}_k)^2$, and the probability that two arbitrary annotators, making arbitrary choices, will happen to agree on any category is the sum of the above values over all labels.

$$\text{[Fleiss's } \kappa\text{]} \qquad \mathrm{A_e} = \frac{1}{N^2}\sum_k (\mathbf{n}_k)^2 \tag{5}$$

Note that the above formula is a *biased estimator* of the expected agreement in the population from which the reliability sample is drawn.

Krippendorff's α is calculated in a similar fashion using expected disagreement, summing the expected coincidences of disagreeing labels scaled by the appropriate distances. Additionally, α uses the scaling factor $1/N(N-1)$ for an unbiased estimator of the expected disagreement in the population.

$$\text{[Krippendorff's } \alpha\text{]} \qquad \mathrm{D_e} = \frac{1}{N(N-1)}\sum_{k_1}\sum_{k_2} \mathbf{n}_{k_1}\mathbf{n}_{k_2}\mathbf{d}_{k_1 k_2} \tag{6}$$

With the nominal distance metric, expected disagreement of Krippendorff's α (Eq. 6) is nearly identical to the complement of the expected agreement of Fleiss's κ (Eq. 5), the only difference being the scaling factor. When N is large and agreement is reasonably high, the difference between Fleiss's κ and Krippendorff's α is very small.

As mentioned above, the values of κ and α range from -1 to 1, where 1 signifies perfect agreement and 0 denotes an agreement level similar to what would be expected by arbitrary annotation. How to interpret the intermediate values is not very clear, and several scales have been proposed in the literature. The computational linguistic community appears to have settled on the recommendation of Carletta [8], accepting coefficient values above 0.8 as reliable, with somewhat lower values also considered acceptable in certain circumstances (Carletta was following the standards set by Krippendorff [23, page 147]). A detailed discussion of coefficient values can be found in Artstein and Poesio [1, Sect. 4.1.3], but overall the emerging consensus appears reasonable. However, a single value can never capture the complexities of a full annotation task, as different aspects of the annotation will be reliable to varying degrees. The next section looks into more detailed reliability analysis, which is intended to give a nuanced understanding of the reliability of a specific annotation effort.

3 Using the Standard Agreement Measures

Agreement coefficients are convenient as a broad assessment of the reliability of an annotation process. As such, it has become common practice to report the reliability of an annotation effort with an overall agreement score. However, reducing an annotation to a single coefficient value carries the risk that the coefficient only represents certain facets of the annotation, possibly hiding important aspects which are less reliable. For a complex annotation task (and pretty much every linguistic annotation task is complex at some level), it is important to investigate reliability at a finer grain than is provided by an overall agreement coefficient. This section explores several ways to use agreement coefficients to get more nuanced insights into four factors that add complications to a reliability analysis: diversity in the underlying data, similarities between the labels, differences in the difficulty of individual items, and differences between individual annotators and annotator populations.

3.1 Diversity in the Underlying Data

An annotation scheme is intended to apply to a set of underlying data, which may be heterogeneous even when coming from a single source. An example of such heterogeneous data is reported in Artstein et al. [2]: four annotators rated the appropriateness of responses given by an interactive question-answering character, on an integer scale of 1–5 (1 being incoherent, 5 being fully coherent). This is a simple task, all the data come from similar dialogues, and reliability turned out to be fairly high ($\alpha = 0.786$). However, we noticed differences in reliability for distinct types of character utterances, which were interleaved throughout the dialogue. When the character had high confidence that he understood the user question, he attempted to answer it directly, giving an *on-topic response*; but when the character's confidence was low he attempted to evade a direct answer by issuing an *off-topic response*. Broken down by character utterance type, the annotators achieved fairly high reliability on rating the coherence of on-topic responses ($\alpha = 0.794$), but were pretty much at chance level on the off-topics ($\alpha = 0.097$).

The results appear puzzling, because off-topic responses constitute about 51% of the data, yet their low reliability has little effect on the coefficient value of the overall annotation. To see how this result comes about we need to examine the actual annotation pattern. It is difficult to visualize four annotators together, so we will look at just two of the annotators; the pattern is similar with the other pairs. Table 1 shows the ratings of one annotator against another, separating the ratings for the character's off-topic and on-topic responses. We observe that the on-topic responses are anchored at two corners on the diagonal – 195 responses (84%) are either maximally coherent or maximally incoherent according to both annotators; this demonstrates reliability, and accounts for the high value of the agreement coefficient. The off-topic responses are only anchored at one corner and the disagreements fan out from there, providing little evidence that the annotators can discriminate reliably between different levels of coherence. When the tables are superimposed on one another, we once again get

Table 1 Utterance coherence by two of the annotators in Artstein et al. [2]

	Off-topic (N = 242)					On-topic (N = 232)				
	1	2	3	4	5	1	2	3	4	5
1	90	19	20	4		30	1			
2	32	20	12	3		8	1	1	1	1
3	12	3	8			1	1	1	1	2
4	9	5	4	1			2		3	3
5								3	7	165
	$\alpha = 0.137$					$\alpha = 0.936$				
	$\alpha = 0.859$									

a table that's anchored at both corners, which is why reliability for the pooled data is high.

The conclusions we draw from such data are fairly complex. It would be clearly misleading to just look at the pooled labels and conclude that the annotation is reliable as a whole. Instead, the data support the following conclusions. Rating the coherence of on-topic responses is reliable. It is also reliably demonstrated that coherence of off-topic responses is generally low – this conclusion comes from the space occupied by the off-topic responses in the pooled data. However, no conclusions can be drawn about the relative coherence of individual off-topic responses; specifically, we cannot conclude that those off-topic responses that received a higher rating by both annotators are any more coherent than the others – these agreements may well be flukes. Finally, the reliability study supports the conclusion that rating the coherence of direct answers (on-topics) is an easier task than rating the coherence of attempts at answer evasion (off-topics).

> When studying annotation of heterogeneous data, agreement should be calculated and reported for the homogeneous subparts of the data, in addition to the data as a whole.

The possibility of low agreement on subparts of the data but high agreement overall is familiar from correlation studies (for the similarities between agreement and correlation coefficients, see Krippendorff [21]). We could find, for example, that there is no meaningful correlation between age and weight in adult elephants and no meaningful correlation between age and weight in adult mice, but when we pool the two populations together, a very strong and significant correlation emerges, because the elephants are both older and heavier than the mice. If we had two annotators estimate the weight of the animals, we might find that they agree at about chance level when estimating weights of elephants, and likewise agree at about chance level when estimating the weights of mice, but pooling the results together brings agreement up to near-perfect levels, because both annotators estimate substantially

Table 2 Hypothetical agreement on tagging offers

Interrogative			Declarative		
	info-req	offer		offer	assert
info-req	78	8	offer	3	7
offer	12	2	assert	9	81
	$\alpha = 0.058$			$\alpha = 0.187$	
		$\alpha = 0.698$			

higher weights for the elephants than for the mice. Such an example would show that the annotators cannot discriminate between individual elephants nor distinguish between individual mice, but they can clearly differentiate elephants from mice.

The effect is not limited to annotations with numerical values; it can occur in categorical annotations as well. Think of a simple dialogue act tagging scheme intended to identify offers. A syntactic question such as *Would you like some tea?* can be ambiguous between an information request and an offer; similarly, a syntactic declarative such as *You need milk in your tea* can be ambiguous between an offer and an assertion. In a hypothetical reliability study, two annotators classify 100 interrogatives and 100 declaratives as information requests, offers and assertions, with the results in Table 2. When calculated separately for interrogatives and declaratives, reliability is fairly low – only 5 and 18% above chance. But pooled together, reliability jumps up to almost 70%, which is considered quite respectable for linguistic annotations. Of course, it would be wrong to conclude from the pooled results that annotators can reliably identify offers; the high agreement only shows that annotators can reliably distinguish between questions and statements.

3.2 Similarity Between Labels

Reliability varies not only when the data are heterogeneous: even with homogeneous data, reliability can be higher for some distinctions than for others. One of the main reasons for reliability testing is to identify specific distinctions in the annotation scheme which are less reliable, that is specific labels which are easily confused with one another. In some cases, the remedy may be to merge labels in order to arrive at a more robust annotation. When labels need to be conflated, it is generally better to rewrite the annotation guidelines and test them rather than merge the labels post-hoc, because the annotators' choice of labels is influenced by all the options they can choose from.

However, when the label set is designed from the outset with some structure, it may make sense to test reliability at multiple levels at once, since working at multiple levels reflects the process the annotators go through when making their choices. An example for such a label set is given in Artstein et al. [3], which tested the semantic coverage of an authored domain. Three annotators used a hierarchical tool to match

Table 3 Reliability at various levels of tags [3]

	α	D_o	D_e
Fully specified act	0.489	0.455	0.891
Dialogue act type	0.502	0.415	0.834
In/out of domain	0.383	0.259	0.420

user utterances to fully specified dialogue acts, selecting first a dialogue act type, followed by properties of the specific act. For example, the utterance *Okay, where have you seen him?* would be mapped to a dialogue act by first selecting the type "wh-question", then an object "strange_man", and finally an attribute "location".

Okay, where have you seen him?	`wh – question`
	`object: strange_man`
	`attribute: location`

Annotators were instructed to match an utterance to the most appropriate dialogue act available in the domain; if none were appropriate, they marked the utterance as "unknown".

We tested reliability both at the level of fully specified dialogue acts and at the level of dialogue act types. Note that the type level is not equivalent to having annotators mark dialogue act types alone, because even the type-level annotation was tied to the domain. For example, the domain did not include any information about whether the character owned a gun. Consequently, the utterance *Do you own a gun?* did not correspond to any existing fully specified act; it was therefore marked as "unknown" by all annotators even though it clearly fits the type "yes-no question". Given this type of scheme, raw disagreement (not corrected for chance) is necessarily lower on the type-level tags than on the fully specified ones, but the difference was rather small (Table 3). Chance-corrected agreement also didn't differ by much, showing that disagreements were mostly concentrated on dialogue act categorization rather than on the specific content of the utterances.

> When annotation labels have an internal structure, it may be acceptable to calculate agreement on different aspects of the same annotation. This is justified when the different aspects reflect separate and distinct decisions made by the annotators, thus reflecting different facets of a complex annotation process.

We also performed a transformation, conflating all the specified dialogue acts into one, and contrasting that with "unknown". This makes a binary distinction of whether the utterance's meaning is close enough to a representation that exists in a domain. While not strictly part of the hierarchy, such conflation is justified because the decision on whether or not to consider an utterance as fitting into the annotation scheme is one that is made by the annotators for each individual instance, at least implicitly. Raw disagreement is again necessarily lower than either fully specified dialogue acts or dialogue act types, though it turned out to be surprisingly high – 0.259, meaning that on 38.8% of the utterances, one annotator disagreed with the others on whether or not the utterance fits the domain (for 3 annotators performing a binary distinction, each item is either in full agreement or a 2–1 split). Table 3 also shows that chance-corrected agreement on the derived binary distinction was lower than on the original task. Since chance-corrected agreement measures annotators' ability to discriminate between categories, we conclude that the task of determining whether an utterance fits the specific domain is a fairly difficult one, probably because the criteria for what constitutes a good fit were not defined clearly. A follow-up study [4] showed that expanding the set of fully specified dialogue acts increases domain coverage on held-out data, but the reliability of the in/out decision did not increase with a wider domain, suggesting that whatever the content is that is covered by the representation scheme, the boundary between what is covered and what is not remains fuzzy.

3.3 Items of Varying Difficulty

Another source of variation, beyond data heterogeneity and label similarity, is variation in the inherent difficulty of individual items: some items are more difficult than others because they are not characterized well by the scheme's category labels, or they lie close to a boundary between labels, or are inherently ambiguous. Identifying difficulty with individual items typically requires more than two annotators, to distinguish cases of genuine difficulty from simple errors. In a study on referential ambiguity, Poesio et al. [29] used 18 annotators working on a single text; while annotators were able to mark items explicitly as ambiguous, many more items were implicitly identified as ambiguous through systematic disagreements between annotators. A different approach was used by Passonneau et al. [28] to infer item-level difficulty in a task of word sense annotation, using 6 trained annotators and 14 crowdsource annotators. Rather than infer difficulty directly from the disagreements on individual items, a graphical model is learned with latent parameters for instance difficulties, true labels, and annotator accuracies; item difficulty is then read from the model. The use of graphical models to learn from annotator discrepancies will be explored further in Sect. 4.

> To identify the extent of individual item difficulty, it is recommended to conduct a reliability study with multiple annotators.

Variation in difficulty does not necessarily show up at the level of individual items; it can also come from broader differences in the source data. Kang et al. [19] calculated the reliability of identifying head nods and smiles in video using two annotators, achieving overall reliability of $\alpha = 0.60$ for head nods and $\alpha = 0.66$ for smiles. In this task the notion of instance difficulty is not very well defined –

agreement was calculated on 50-millisecond time slices, and adjacent instances often received identical labels because head nods and smiles typically last for much longer than 50ms. Differences were noted at the level of individual video clips, where there was substantial variation in reliability (each clip depicted a different person). For head nods, α ranged from -0.16 to 0.99, with agreement on some clips lower than expected by chance; for smiles, α ranged from 0.17 to 0.98 (chance correction for individual clips was always performed using the expected agreement derived from the pooled annotation data). This variation in reliability probably indicates variation in difficulty of the individual video clips – that is, that smiles and head nods are harder to detect on some people than others.

3.4 Differences Among Annotators

One further source of variation in reliability is the annotators. An underlying assumption behind annotation efforts is that individual annotators are roughly equivalent. Krippendorff [24] explicitly builds the requirement that annotators be interchangeable into the definition of α, insisting that all the knowledge required for the annotation task be derived from the written manuals. Annotator interchangeability is an ideal, which might be workable to some extent for very simple annotation tasks. But practical experience with linguistic annotation shows that there are differences both between annotator populations and between individual annotators.

Annotators used in linguistic efforts often have some linguistic training, partly due to the population that is available for recruitment (linguistics students), and partly because it is believed that linguistic training makes better annotators, as shown for example by Kilgarriff [20] for word-sense annotation. Linguistic expertise, however, is not the only relevant dimension along which annotators differ. Scott et al. [32] found systematic differences based on medical expertise when annotators rated hedges in medical records as likelihoods: for each hedge (for example "*possible* early pneumonia" or "*could* represent pneumonia"), annotators were asked to judge how the doctor who wrote the hedge viewed the likelihood of the indicated medical condition. The results showed that annotators with medical training tended to judge each hedge as expressing a greater likelihood than the corresponding judgments by annotators without medical training: that is to say, when a doctor reads a statement like "possible early pneumonia", written by another doctor, she would interpret the statement as expressing greater likelihood than a lay person would. Since these medical records are written by doctors and for doctors, it is reasonable to assume that in this case, the doctors' interpretation is a better reflection of the writers' intention.

Even when annotators reflect a homogeneous background, there may still be substantial variability between them. And while confidence intervals for agreement coefficients can be estimated through resampling of the annotated items [16], this method cannot be used to quantify annotator variation, because resampling annotators would result in measuring agreement between an annotator and herself. Nevertheless, useful insight can be gained by simply measuring agreement for all the subgroups (pairs, triples) of annotators that participated in the reliability study. Passonneau et al. [26]

Table 4 Agreement among subgroups of annotators [11]

α	Annotators	α	Annotators	α	Annotators	α	Annotators
0.593	B E	0.680	B C D E	0.715	A C D E	0.740	A B C D
0.617	D E	0.689	A B D E	0.721	B C D	0.754	A D
0.646	B D E	0.696	A D E	0.723	A B	0.754	A B C
0.656	C E	0.697	A E	0.727	A C E	0.759	A C D
0.670	B C E	0.702	B D	0.727	A B D	0.801	A C
0.673	C D E	0.708	A B C D E	0.727	C D		
0.678	A B E	0.709	A B C E	0.737	B C		

calculate reliability for subsets of annotators in order to identify maximal groups that have high internal agreement; they show that in some cases, dropping just a few annotators can result in very good agreement among the remaining annotators. Results from a different experiment are shown in Table 4, with agreement among five annotators who judged the adequacy of a Natural Language Generation output relative to the semantic representation that served as input [11]. It is apparent from the table that annotator E is somewhat of an outlier, who tends to disagree with the other annotators more than they disagree among themselves (this does not mean that annotator E is worse than the others, but this difference should be investigated further). Among the other annotators there is no clear outlier, yet chance-corrected agreement varies by almost 10%, from $\alpha = 0.702$ for annotators B and D to $\alpha = 0.801$ between annotators A and C. Looking at agreement values for the different groups of annotators can give a better sense of how stable the agreement value is for a particular annotation effort.

In a reliability study with more than two annotators, differences between the annotators should be investigated by calculating agreement among subgroups of annotators.

3.5 Summary

The examples in this section have demonstrated one of the major pitfalls of using agreement coefficients, namely the fact that a single coefficient value can mask complex patterns in an annotation effort. Annotated corpora can be reliable in some parts but not others, or reliable in some aspects but not others, and detailed measurements can help identify the extent to which the various parts or aspects of an annotation can be trusted. Specific sources of variation within a single annotation effort include diversity in the underlying data, similarities between the labels, differences in the difficulty of individual items, and differences between individual annotators. Agreement measures are a useful tool for studying this variation, and I therefore advocate for conducting and reporting detailed analyses rather than just an overall coefficient value. These detailed analyses include separate agreement calculations for homogeneous subparts of the data; separate analysis of different aspects of a complex annotation task; using multiple annotators to uncover difficulty in individual items; and calculating agreement on subgroups of annotators to uncover systematic differences between the annotators themselves.

4 Exploiting Annotator Disagreement

The previous section has shown how agreement coefficients can be used to extract insight about the annotation process and assess various aspects of annotation reliability. The detailed analyses can uncover unreliable facets of an otherwise reliable annotation process, and the underlying methodology assumed so far has been that of the textbook use case – quantify agreement in order to improve annotation guidelines and arrive at a reliable process. However, the goal of developing a process that is sufficiently reliable in all the relevant aspects is not always attainable. When annotation is not reliable (or not reliable enough), it is still possible to exploit this lack of reliability – the disagreements between the annotators – in order to make use of the annotations for linguistic applications.

Unlike fields like content analysis, where inferences are drawn directly from annotated data, the use of annotations in computational linguistics is typically indirect: annotated data are used for training computational processes via machine learning, and it is these processes and their outputs that are of interest. Reidsma and Carletta [30] show that annotation reliability does not imply that the annotated data are suitable for machine learning. This is because machine learning is sensitive not only to the amount of noise in the training data, but also (and more importantly) to its location. Reidsma and Carletta present a series of experiments that show successful machine learning with high levels of noise (and hence low annotation reliability), when noise is distributed uniformly; contrasted with unsuccessful machine learning with lower levels of noise (and hence higher annotation reliability), when noise is localized in a way that interferes with machine learning. Hence, goes the argument, annotation reliability is neither necessary nor sufficient for successful machine learning, and thus it is not important for linguistic annotation. Unfortunately, the reported experiments were conducted using synthetic noise. This only demonstrates that a dissociation between annotation reliability and machine learning success is a theoretical possibility; a dissociation has not been shown to occur in actual annotation tasks.

Recent research by Passonneau and Carpenter [27] shows that given sufficient redundant data, correct labels can be recovered from noisy and unreliable annotations using statistical methods. Annotations by multiple annotators are used to learn a graphical annotation model which infers the correct labels from the annotators' labels. The model parameters include a true label for each instance, a probability distribution of the true labels, and for each annotator and true label, a probability distribution of observed labels assigned by the annotator to instances of the true label; the latter set of parameters reflects biases of individual annotators and tendencies for confusion among labels. The model parameters are learned through maximum likelihood estimation, and the resulting annotation model corrects for many of the errors made by the annotators themselves. Unlike the model of Passonneau et al. [28] described in Sect. 3.3, the Passonneau and Carpenter [27] model does not include parameters for instance difficulty; however, the model provides a probability for the inferred label for each instance, giving an estimate of the quality (or confidence) for each individual label. Passonneau and Carpenter also show that as a practical matter, despite the fact that building an annotation model requires more data per instance than traditional annotation, acquiring such data through crowdsourcing can be done faster and at a lower total cost.

A similar graphical model is presented by Hovy et al. [17]. The parameters of this model are a true label for each instance, and for each annotator, a trustworthiness score (the probability of making an informed judgment resulting in the true label) and a probability distribution of labels when making an uninformed judgment (which could also result in the true label by mere chance). Unlike the model of Passonneau and Carpenter [27], this model does not capture relations between true label and annotator output: when the annotator is acting in an untrustworthy manner, the output is independent of the true label.

The ultimate purpose of developing reliable annotation processes is to arrive at a set of correct labels; therefore, the ability to derive correct labels from unreliable annotation appears to obviate the need for reliable annotation. However, it is not clear whether learning a model from unreliable annotations gives results that are comparable to traditional trained annotators. Passonneau and Carpenter [27] show that the learned annotation model results in a different distribution of labels than that of the trained annotators; the claim that the learned model is better is based primarily on the observation that the trained annotator labels are more similar to the crowdsource plurality than to the learned model, which is considered better than the plurality vote. Additionally, the model labels come with confidence scores, which the trained annotator labels lack. However, since learning an annotation model requires many annotations for each instance, it is only feasible for tasks which can be designed to be performed with minimal instruction and training.

5 Conclusion

Linguistic annotation is used for tasks that cannot be performed mechanically, and whenever human judgments are called for, there will be some variation. In order to make use of annotated data, it is important to know what variation exists in the data, and to assess how this variation affects the intended use. Having multiple annotators work on at least a portion of the data is essential for an estimate of the amount of variation, and formal agreement measures are useful for quantifying the variation. Appropriate measures of inter-annotator agreement can help assess the reliability of an annotation process, but this has to be done with care, because reliability is complex, affecting different aspects of the annotation to varying degrees.

It is therefore important to conduct detailed investigations into each annotation effort, along the various dimensions in which annotation reliability can vary. When sufficient annotations are available, it is also possible to exploit the variation among annotators and use machine learning to infer the correct labels. In either case, publications should report relevant results of the detailed agreement studies, rather than just a blanket statement about overall reliability.

Acknowledgements I thank the editors of this volume and two anonymous reviewers for valuable feedback and comments on an earlier draft. Any remaining errors and omissions are my own.

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
2. Artstein, R., Gandhe, S., Gerten, J., Leuski, A., Traum, D.: Semi-formal evaluation of conversational characters. In: Grumberg, O., Kaminski, M., Katz, S., Wintner, S. (eds) Languages: From formal to natural. Essays dedicated to Nissim Francez on the occasion of his 65th birthday, Lecture Notes in Computer Science, vol. 5533, pp 22–35. Springer, Heidelberg (2009)
3. Artstein, R., Gandhe, S., Rushforth, M., Traum, D.: Viability of a simple dialogue act scheme for a tactical questioning dialogue system. DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue, pp. 43–50. Stockholm, Sweden (2009)
4. Artstein, R., Rushforth, M., Gandhe, S., Traum, D., Donigian, A.: Limits of simple dialogue acts for tactical questioning dialogues. In: Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, pp. 1–8. Barcelona, Spain (2011)
5. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? A meta-analytic investigation. Comput. Linguist. **37**(4), 699–725 (2011)
6. Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications through limited questioning. Public Opin. Q. **18**(3), 303–308 (1954)
7. Byrt, T., Bishop, J., Carlin, J.B.: Bias, prevalence and kappa. J. Clin. Epidemiol. **46**(5), 423–429 (1993)
8. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
10. Craggs, R., McGee Wood, M.: Evaluating discourse and dialogue coding schemes. Comput. Linguist. **31**(3), 289–295 (2005)
11. DeVault, D., Traum, D., Artstein, R.: Making grammar-based generation easier to deploy in dialogue systems. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics, pp. 198–207. Columbus, Ohio, http://www.aclweb.org/anthology/W/W08/W08-0130 (2008)
12. Di Eugenio, B., Glass, M.: The kappa statistic: a second look. Computational Linguistics **30**(1), 95–101 (2004)
13. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971)

14. Fleiss, J.L.: Measuring agreement between two judges on the presence or absence of a trait. Biometrics **31**(3), 651–659 (1975)
15. Fort, K., François, C., Galibert, O., Ghribi, M.: Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 1474–1480. Istanbul, Turkey (2012)
16. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. **1**(1), 77–89 (2007)
17. Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: Learning whom to trust with MACE. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 1120–1130. Atlanta, Georgia, http://www.aclweb.org/anthology/N13-1132 (2013)
18. Hsu, L.M., Field, R.: Interrater agreement measures: comments on kappa$_n$, Cohen's kappa, Scott's π, and Aickin's α. Underst. Stat. **2**(3), 205–219 (2003)
19. Kang, S.H., Gratch, J., Sidner, C., Artstein, R., Huang, L., Morency, L.P.: Towards building a virtual counselor: Modeling nonverbal behavior during intimate self-disclosure. In: Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Valencia, Spain (2012)
20. Kilgarriff, A.: 95% replicability for manual word sense tagging. In: Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 277–278 (1999)
21. Krippendorff, K.: Bivariate agreement coefficients for reliability of data. Soc. Methodol. **2**, 139–150 (1970)
22. Krippendorff K (1978) Reliability of binary attribute data. Biometrics 34(1):142–144, letter to the editor, with a reply by Joseph L. Fleiss
23. Krippendorff, K.: Content analysis: an introduction to its methodology. Sage, Beverly Hills, CA, chap **12**, 129–154 (1980)
24. Krippendorff, K.: Content analysis: an introduction to its methodology. 2nd edn. Sage, Thousand Oaks, CA, chap **11**, 211–256 (2004)
25. Krippendorff, K.: Reliability in content analysis: some common misconceptions and recommendations. Hum. Commun. Res. **30**(3), 411–433 (2004)
26. Passonneau, R., Habash, N., Rambow, O.: Inter-annotator agreement on a multilingual semantic annotation task. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pp. 1951–1956. Genoa, Italy, http://www.lrec-conf.org/proceedings/lrec2006/summaries/634.html (2006)
27. Passonneau, R.J., Carpenter, B.: The benefits of a model of annotation. Trans. Assoc. Comput.l Linguist. **2**, 311–326, http://www.aclweb.org/anthology/Q/Q14/Q14-1025.pdf (2014)
28. Passonneau, R.J., Bhardwaj, V., Salleb-Aouissi, A., Ide, N.: Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. Lang. Res. Eval. **46**(2), 219–252 (2012)
29. Poesio, M., Sturt, P., Artstein, R., Filik, R.: Underspecification and anaphora: theoretical issues and preliminary evidence. Discourse Processes **42**(2), 157–175 (2006)
30. Reidsma, D., Carletta, J.: Reliability measurement without limits. Comput. Linguist. **34**(3), 319–326 (2008)
31. Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Q. **19**(3), 321–325 (1955)
32. Scott, D., Barone, R., Koeling, R.: Corpus annotation as a scientific task. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 1481–1485. Istanbul, Turkey (2012)
33. Zwick, R.: Another look at interrater agreement. Psychological Bulletin **103**(3), 374–378 (1988)