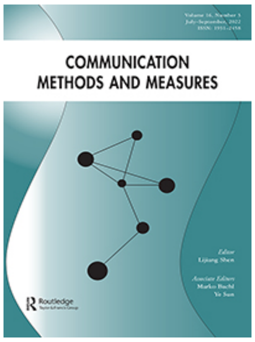


Communication Methods and Measures

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue

Nathan Duran, Steve Battle & Jim Smith

To cite this article: Nathan Duran, Steve Battle & Jim Smith (2022) Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue, Communication Methods and Measures, 16:3, 182-214, DOI: 10.1080/19312458.2021.2020229

To link to this article: https://doi.org/10.1080/19312458.2021.2020229

Published online: 17 Jan 2022.

COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 3, 182–214
https://doi.org/10.1080/19312458.2021.2020229

Routledge
Taylor & Francis Group

Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue

Nathan Duran, Steve Battle, and Jim Smith

Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK

ABSTRACT

We present the Conversation Analysis Modeling Schema (CAMS), a novel dialogue labeling schema that combines the Conversation Analysis concept of Adjacency Pairs, with Dialogue Acts. The aim is to capture both the semantic and syntactic structure of dialogue, in a format that is independent of the domain or topic, and which facilitates the computational modeling of dialogue. A labeling task undertaken by novice annotators is used to evaluate its efficacy on a selection of task-oriented and non-task-oriented dialogs, and to measure inter-annotator agreement. To deepen the "human-factors" analysis we also record and examine users' self-reported confidence scores and average utterance annotation times. Inter-annotator agreement is shown to be higher for task-oriented dialogs than non-task-oriented, though the structure of the dialogue itself has a more significant impact. We further examine the assumptions around expected agreement for two weighted agreement coefficients, Alpha and Beta, and show that annotators assign labels using similar probability distributions, small variations can result in large differences in agreement values between biased and unbiased measures.

Human conversational interactions are, naturally, a complex phenomenon. When we take part in such interactions, we may utilize a range of visual, verbal, and linguistic cues to interpret the intentions of other participants, formulate responses and organize turns of talk (Goodwin, 1981). Even when considered solely in an audio or text-based form, the utterances of an interaction cannot be fully understood on an individual basis, but rather must be interpreted within the context of their position within the sequence of utterances (Ekman & Scherer, 1984). The question of how such intricate conversational data can be represented in a computationally practical format remains an open problem within Natural Language Processing (NLP) research.

The predominant approach to representing dialogue semantics, for the purpose of NLP, is the use of Dialogue Acts (DA). Originating from John Austin's "illocutionary act" theory (Austin, 1962), and later developed with John Searle's "speech acts" (Searle, 1969), a DA defines the semantic content and communicative function of a single utterance of dialogue, for example, a question, statement or greeting. The utility of DA, as a set of labels for a semantic interpretation of a given utterance, has led to their use in many NLP applications. In dialogue management systems they have been used as a representation of user and system dialogue turns, as a set of possible system actions, and as a means of dialogue state tracking (DST) (Cuayahuitl et al., 2016; Firdaus et al., 2020; Ge & Xu, 2015; Griol et al., 2008; Keizer & Rieser, 2017; Li et al., 2017). For spoken language translation Kumar et al. (2008) utilized the contextual information provided by DAs to improve accuracy in phrase-based statistical speech

CONTACT Nathan Duran nathan.duran@uwe.ac.Uk Department of Computer Science and Creative Technologies, University of the West of England, Frenchay Campus, Coldharbour Ln, Bristol BS16 1QY, UK

This article has been republished with minor changes. These changes do not impact the academic content of the article.

translation. They have also been used to analyze the structure of dialogue within the intelligent tutoring domain (Boyer et al., 2009, 2010), and everyday conversations (Iseki, 2019). While DA do provide valuable semantic and intentional information, they naturally consider utterances as an isolated unit. In so doing, they fail to recognize the sequential nature of interactions, and the influence that both context and position have, on the production and meaning of an utterance (Clift, 2016; Ekman & Scherer, 1984). As Clift (2016), points out, "*the form of an utterance alone cannot necessarily be relied upon to deliver how it is understood by its recipient.*" Consider the use of "Okay" in the following examples. In the first instance speaker B uses "Okay" in response to a question. In the second instance, speaker A uses "Okay" as confirmation that a response has been heard and understood.

```
1  A: How are you?     2  A: Do you need help with that?
   B: Okay                B: No thank you.
                          A: Okay
```

What is needed, then, is a method of representing not just the semantics of single utterances but the context within which they were produced and their contribution to the interaction as a whole. For this, we turn to the study of human conversation. Conversation Analysis (CA) is an area of sociological research that aims to define, and analyze, constructs that facilitate turn-taking in human conversations (Sacks et al., 1974). Some key principles of CA are: that turns of talk have some organizational structure; that the structure itself has a descriptive quality for the utterances produced; and in turn, helps to shape the future utterances of the interaction (Schegloff, 2007; Sidnell, 2010). Within CA, this structure is defined using the concept of the Adjacency Pair (AP) as the base units of sequence-construction in talk. Utterances are labeled with AP such that they describe the relational structure *between* utterances of a dialogue. Therefore, DA labels may be considered descriptions of the *intra-utterance* features of a dialogue, while AP represent the *inter-utterance* features.

In this article, we introduce the Conversation Analysis Modeling Schema (CAMS). With CAMS, we hope to produce richer and more expressive representations of dialogue, in a computationally compatible format, to aid in the development of Conversational Artificial Intelligence (CAI) tasks, such as dialogue management and DST, as well as other NLP applications. The schema defines a domain agnostic annotation scheme for dialogue that is aligned with relevant theories from within the CA literature, to express the general structure of an interaction, while leveraging the descriptive power of the DA for individual utterances. The schema defines both AP and DA labels which combine to form AP-types. The AP-type labels are intended to capture the semantic and syntactic structure of an interaction, in a format that is independent of the domain or topic, and which facilitate the computational modeling of dialogue. We evaluate CAMS by means of an annotation study, calculate measures of inter-annotator agreement in order to assess its efficacy when applied to both task and non-task-oriented dialogs, and determine the extent to which novice annotators arrive at a shared understanding of the categories within the coding scheme. We also record users' self-reported annotation confidence scores, and average utterance annotation times, as an additional human-factors analysis. Through these measures, we hope to evaluate considerations, such as, choice of agreement coefficient, source of dialogue material, and annotator characteristics or behaviors, which may affect application of the schema for further annotation tasks.

The following section provides a full description of CAMS, its labels and annotation guidelines. Then, Inter-Annotator Agreement measures are outlined, and the distance functions used for weighted agreement coefficients within this study are defined in Weighted Coefficient Distance Functions section. Data and Methods gives details of the methodological setup, selection of participants and dialogue corpora, before discussing the results obtained from the annotation procedure in Results and Discussion. And finally, our Conclusions are drawn.

Conversation analysis modeling schema overview

CAMS is intended to combine concepts of DA and AP into a single annotation scheme that is able to capture the semantic and syntactic structure of a dialogue at the *inter* and *intra* utterance level. Additionally, AP and DA may be applied to any type of conversational interaction, independent of domain and topic, and as such, the schema is entirely domain agnostic and applicable both to task and non-task-oriented dialogs.

The schema defines two sets of labels, DA and AP, which are combined to form AP-type labels. When applying the schema, the intent is to assign each utterance of a dialogue one DA and one AP label, which together are considered the AP-type label for that utterance. The AP-type labels, for a fully annotated dialogue, can then be viewed as a representation of its semantic and syntactic structure, as described above. It should be noted that the concept of a *typed AP* is a key feature of AP present within the CA literature (Clift, 2016; Liddicoat, 2007; Schegloff, 2007; Sidnell, 2010). However, the standard annotation schemes for CA do not strictly require each utterance of dialogue to be labeled with an AP. Additionally, CA annotation often includes non-verbal sounds, pauses and other types of disfluencies. Gaps in annotations, where utterances are not labeled with AP, and other forms of non-verbal annotation, for example, "breathing," are generally undesirable for computational purposes. CAMS, therefore, is an attempt to define these concepts, and how they may be applied, into a computationally compatible format where each utterance is labeled with an AP-type. The following sections provide an overview of AP, DA, and AP-types, and their respective sets of labels defined within the schema.[1]

Adjacency pairs

AP are the base units of sequence-construction in talk, and in their basic unexpanded form, comprise of two turns by different speakers that take place one after the other. The initial turn is called the *First Pair Part* (FPP) and initiates an exchange, the second turn is a *Second Pair Part* (SPP) which is responsive to the prior FPP. AP may also be "type related," for example, a question and an answer (Schegloff, 2007). This *pair-type* relation has the useful property of limiting the range of possible SPP responses to a given FPP, for example, a question could be followed by an answer (though not necessarily) but is unlikely to be followed by a greeting (Liddicoat, 2007). For the purpose of analysis within NLP, and particularly dialogue systems, this is advantageous because it reduces the set of all possible SPP responses to just a few types. Participants in conversation orient to this basic sequence structure in developing their talk and set up expectations about how talk will proceed. Within the schema they are assigned the FPP-base and SPP-base labels, and these represent the core activity through which speakers accomplish their communicative goals, or actions.

```
A: What time is it?  FPP-base

B: Three o' clock.   SPP-base
```

Expansions

To account for more complex dialogue structures, AP also include the concept of *expansion*, which allows the construction of sequences of talk that are made up of more than one AP, while still contributing to the same basic action (Liddicoat, 2007). Sequence expansion is constructed in relation to a base sequence of a FPP and SPP in which the core action under way is achieved. There are three types of expansion pairs *Pre, Post*, and *Insert*.

Pre-expansions. Are designed to be preliminary to some projected base sequence and may be considered as preludes to some other action.

[1] Full label definitions and annotation guidelines are available at: nathanduran.github.io/Conversation-Analysis-Modeling-Schema

A:	What you doing?	*FPP-pre*
B:	Not much.	*SPP-pre*
A:	Wanna drink?	*FPP-base*
B:	Sure.	*SPP-base*

Post-expansions. Allow talk to occur after a base sequence, which is recognizably associated with the preceding sequence.

A:	What is the weather like today?	*FPP-base*
B:	Forecast for cloudy skies today.	*SPP-base*
A:	Okay.	*FPP-post*
B:	No problem.	*SPP-post*

Insert-expansions. Occur between base adjacency pairs and separates the FPP and SPP. Insert-expansions interrupt the activity previously underway but are still relevant to that action and allows the second speaker (who must produce the base SPP), to do interactional work relevant to the base SPP. Once the sequence is completed, the base SPP once again becomes relevant as the next action. For example, a question (FPP-base) could be followed by a question (FPP-insert), to elicit information required to better answer the initial question. The insert-expansion is then concluded before completing the original base pair, as in the following example.

A:	Do you know the directions to the zoo?	*FPP-base*
B:	Are you driving or walking?	*FPP-insert*
A:	Walking.	*SPP-insert*
B:	Get on the subway . . .	*SPP-base*

Minimal-Expansions

Because dialogue does not always contain even numbers of utterances, there are also single-utterance *minimal-expansions*, for utterances that do not belong to conventional AP. CAMS defines three types of minimal-expansion *Pre, Post*, and *Insert*, which behave in a similar manner to their expansion counterparts. That is, they must be produced before, after, or inside a base sequence. These are closely related to the idea of minimal post-expansions (Schegloff, 2007), in that they are not designed to project any further sequences of talk, but rather open, close or add to sequences respectively. The primary role is to allow for additional turns that behave as expansions but consist only of one turn. There is no restriction on speaker order for minimal-expansions, which allows the same speaker to produce more than one utterance of different types in succession, or for a speaker to produce one utterance that does not belong to (initiate or conclude) an AP.

A:	When is my dentist appointment?	*FPP-base*
B:	The appointment is at 11 am with your aunt.	*SPP-base*
A:	Thanks.	*Post*

In summary, there are 11 AP in the schema and the set includes: Two labels for the base pair, FPP-base and SPP-base. Six labels for expansion pairs. That is, FPP and SPP for pre, post and insert expansions, as described by Liddicoat (2007) and Sidnell (2010). And three labels for minimal expansions, pre, post, and insert.

Table 1. The CAMS DA labels derived from DiAML and grouped by communicative function.

Communicative Function	DA Labels
Information-seeking	setQuestion, choiceQuestion, propQuestion, checkQuestion
Information-providing	answer, inform, correction
Commissive	offer
Directive	suggest, request
Feedback Positive	accept, conditionalAccept, agree, confirm, feedbackPos
Feedback Negative	decline, disagree, disconfirm, feedbackNeg
Time and Communication	stalling, retraction
Social Management	greeting, goodbye, thanking, acceptThanking, apology, acceptApology

Dialogue acts

Though it was philosophers such as Austin (1962) and Searle (1969), who reconceptualized speech as "actions," the term *dialogue act* was introduced by Bunt (1978). Bunt (2000), argued that a notion of *communicative functions* is required, which establish semantic definitions in terms of dialogue context changes, and further that, communication has many "dimensions" that a speaker can address simultaneously. For instance, "Yes, but what is it?", indicates both an understanding of what was previously said, and a request for more information. From this example we can define DA in terms of two components: i) its *communicative function*, what the speaker is trying to achieve, and ii) the *semantic content*, which describes the information that is being addressed – the entities, their properties, and relations that are referred to. Thus, while DA labels are intended for single utterances of dialogue, they can be both multidimensional (have more than one function), and be prospective, or reactive, to surrounding utterances; a property that is particularly advantageous when viewed in conjunction with the broader structural descriptions provided by AP.

As previously discussed, DA are commonly used for NLP purposes. However, historically there has been quite a range of different labeling schemes developed. Most notably, the Discourse Annotation and Mark-up System of labeling (DAMSL) (Allen & Core, 1997), which was used to annotate the Switchboard Dialogue Act dataset (Jurafsky et al., 1997), and a slight variation was used to label the Meeting Recorder Dialogue Act (MRDA) corpus (Shriberg et al., 2004). Also, corpora created for the development of dialogue systems, such as the Dialogue State Tracking Challenge (DSTC) (Williams et al., 2016), and FRAMES (Asri et al., 2017), typically define their own bespoke set of DA labels. While there is some commonality between them, the net result is a collection of different DA labeling schemes that are, to some degree, incompatible. In a move to address this problem the Dialogue Act Mark-up Language (DiAML) was developed and forms part of ISO 24617 (British Standards Institution, 2012). DiAML was developed as an empirically and theoretically well founded, application independent, DA annotation scheme and is also intended to be used by both human annotators and automatic annotation methods. There seems to be some growing recognition, within the DA research community, of the utility of a standardized method of DA annotation with several attempts to map existing DA labeled corpora to the DiAML scheme (Chowdhury et al., 2016; Mezza et al., 2018). As such, the 27 DA labels defined within CAMS are entirely derived from a subset of DiAML labels. As shown in Table 1, they remain grouped by their communicative function: Information-seeking, information-providing, commissives, directives, feedback, time management, owner and partner communication management, and social obligations management. Note that, within DiAML, the labels *autoPositive* and *autoNegative* represent positive or negative understanding of the previous utterance, for example, "Okay," or "What?." Within CAMS we have converted these into the slightly more intuitive labels of *feedbackPos* and *feedbackNeg*.

Adjacency pair types

In CAMS, an AP-type is simply the product of one AP label, and one DA label, for an utterance of dialogue. The combination of these two labels is considered an AP-type label. Due to the large number of possible combinations, and to allow flexibility, the schema does not explicitly define all valid DA and AP combinations. Instead, annotators should consider the meaning and context within which the individual labels being applied produce AP-types. The following shows a previous example, now fully labeled with both AP and DA, to create AP-types. In the example, *propQ* (propositionalQuestion) is a question that implies, but does not necessitate, a "yes" or "no" answer, and a *choiceQ* (choiceQuestion) where the speaker provides a list of alternatives with the assumption that the addressee knows which one is true, or will select one. The alternative question-type labels are: *setQuestion*, which corresponds to what is commonly termed a "WH-question" in the linguistic literature, that is, questions that typically begin with words such as, "Who," "What" or "How"; and *checkQuestion*, which is produced by the speaker in order to know whether a proposition is true.

A:	Do you know the directions to the zoo?	*FPP-base – propQuestion*
B:	Are you driving or walking?	*FPP-insert – choiceQuestion*
A:	Walking.	*SPP-insert – answer*
B:	Get on the subway . . .	*SPP-base – answer*

Inter-Annotator agreement

Inter-annotator agreement measures can be used as a means of assessing the *reproducibility* of a coding scheme or determining the *reliability* of a produced "gold standard" labeled dataset. Given that the focus of this study is the labeling schema itself, the purpose of measuring inter-annotator agreement refers to the former. That is, determining if the schema is inherently learnable, that the labels applied to utterances are not entirely dependent on the biases of an individual annotator, and that there is a common understanding of the meaning of labels and the utterances to which they are applicable (Craggs & Wood, 2005). It should be noted, that reproducibility is a natural prerequisite to demonstrating reliability of a coding scheme. If annotators produce similar results, they likely have a similar understanding of the annotation scheme and guidelines, and that these are able to represent the desired characteristics of the data (Artstein & Poesio, 2008). Within the literature chance-corrected coefficients, that is, accounting for the probability that annotators select the same label by chance, such as Cohen's Kappa (Cohen, 1960), or Scott's Pi (Scott, 1955), are the preferable measures of inter-annotator agreement (Carletta, 1996; Craggs & Wood, 2005; Di Eugenio, 2000). However, weighted coefficients, such as Krippendorff's Alpha (Krippendorff, 2004), are more suitable to annotation tasks such as this, which require an element of semantic interpretation.

Weighted agreement coefficients

For some annotation tasks it does not make sense to treat all disagreements equally. For example, the DA *choiceQuestion* and *checkQuestion* are semantically more similar than *request* and *accept*. Both Pi and Kappa are limited in such circumstances because they only consider identical labels for agreement. This can result in very poor agreement values and as such they are not considered an acceptable measure of agreement for DA labeling tasks (Artstein & Poesio, 2005b; Geertzen & Bunt, 2010). A solution to this problem is the use of weighted agreement coefficients, which consider the magnitude of disagreement between assigned labels. Cohen (1968), proposed a weighted variation of Kappa for two annotators. More frequently used however, and appropriate for this study, is Krippendorff's Alpha (Krippendorff, 2004), and the Beta statistic, proposed by Artstein and Poesio (2005b). Figure 1 summarizes some of the characteristics of each coefficient with respect to three different dimensions, bias and unbiased (Kappa and Pi), two or multiple coders (multi-Kappa and multi-Pi), and weighted (Alpha and Beta).

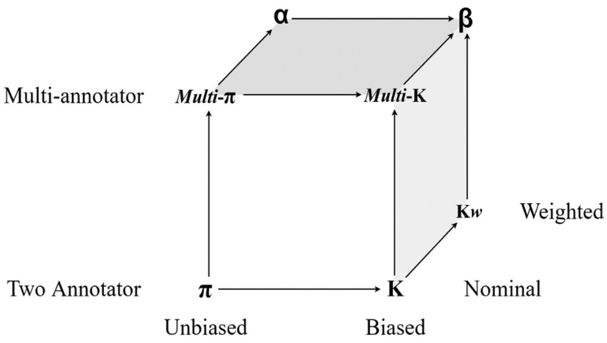


Figure 1. Agreement coefficients in three dimensions, bias, number of coders, and weighted. Adapted from the "Coefficient Cube" (Artstein & Poesio, 2005b).

Both Alpha and Beta are calculated from the observed and expected *disagreements*, rather than the agreement of the previously discussed coefficients. The ratio of observed (o) and expected (e) disagreement is then subtracted from 1 to produce the final agreement value:

$$\alpha, \beta = 1 - \frac{D_o}{D_e} \tag{1}$$

Further, weighted coefficients use a distance function (see section Weighted Coefficient Distance Functions), which returns a value in the range [0, 1] representing the similarity between an arbitrary pair of labels. 0 indicates the two labels are identical and 1 indicates they are completely dissimilar. This value is then used to weight pairs of assigned labels, penalizing those that are more dissimilar. The amount of disagreement for a given item is, therefore, the mean of the distances between all pairwise assignments for that item. The number of annotators who label item i, with label l, is n_{il}. For every label pair l_j and l_k, there are n_{il_j} n_{il_k} pairs of assigned labels for an item, and each has a distance (**d**) of $d_{l_j l_k}$, calculated by the distance function. The mean disagreement for an item is then the sum of all weighted label pairs, divided by the total number of annotator pairs, $a(a-1)$:

$$disagr_i = \frac{1}{a(a-1)} \sum_{j=1}^{l} \sum_{k=1}^{l} n_{il_j} n_{il_k} d_{l_j l_k} \tag{2}$$

Observed disagreement is then the mean disagreement for all items:

$$D_o = \frac{1}{i} \sum_{i \in I} disagr_i \tag{3}$$

Where Alpha and Beta differ, is in their estimations of the distribution of assigned labels for an annotator operating only by chance, that is, how $P(l|a_k)$ is estimated. When calculating D_e, Alpha estimates disagreement on the basis that each annotator assigns labels with the same distribution and therefore considered an *unbiased* coefficient, whereas Beta is *biased*, in that it calculates D_e from the observed distribution of individual annotators.

Alpha

Given the single probability distribution for all annotators, the probability of assigning a label to an item is the number of assignments of the label by all annotators, divided by the total number of assignments – items **i** multiplied by the number of annotators **a**.

$$P(l) = \frac{n_l}{ai} \tag{4}$$

Again, the probability that two annotators assign labels l_j and l_k, is the joint probability of each annotator assigning the label independently. The expected disagreement is, therefore, the sum of the weighted joint probabilities for all label pairs, divided by the total number of assignments:

$$D_e^{\alpha} = \frac{1}{ai(ai-1)} \sum_{j=1}^{l} \sum_{k=1}^{l} n_{l_j} n_{l_k} d_{l_j l_k} \tag{5}$$

Beta

The Beta coefficient is, in essence, multi-annotator generalization of Cohens weighted Kappa (Artstein & Poesio, 2005b); in that, it is a weighted coefficient which considers individual annotators label distributions (bias) and is applicable to more than two annotators. The probability that annotator a, assigns label l, to an item, is the total number of such assignments n_{al}, divided by the total number of assignments for that annotator (the same as Kappa and Multi-kappa):

$$P(l|a_j) = \frac{n_{a_j l}}{i} \tag{6}$$

The probability that two annotators a_m and a_n, selecting different labels l_j and l_k, is $P(l_j|a_m)P(l_k|a_n) + P(l_k|a_m)P(l_j|a_n)$. The probability that a given pair of coders assigns labels l_m and l_n, is the mean of the probabilities for all annotator pairs:

$$P(l_j, l_k) = \frac{1}{ia(ia-1)} \sum_{m=1}^{a-1} \sum_{n=1}^{a} n_{a_m l_j} n_{a_n l_k} + n_{a_m l_k} n_{a_n l_j} \tag{7}$$

The expected agreement for Beta is then, the mean of the probabilities for each pair of labels weighted by the distances:

$$D_e^{\beta} = \sum_{j=1}^{L-1} \sum_{k=j+1}^{L} P(l_j, l_k) d_{l_j l_k} \tag{8}$$

It is worth noting, that if all disagreements are considered equal, with distance 1, then Alpha and Beta produce the same result as their non-weighted equivalents Multi-pi and Multi-kappa. Similarly, if data from only two annotators is used, and the distances are equal, the results are the same as the non-weighted two annotator variants Pi and Kappa.

Weighted coefficient distance functions

The calculation of Alpha and Beta requires a distance function **d**, that returns a distance value in the range [0, 1] for each possible label pair. The value indicates the amount of dissimilarity between the two labels, with 0 indicating they are identical and 1 indicating they are completely dissimilar. In this section 3 distance functions are defined, one for each of the label types defined within the schema. The constraints suggested by Artstein and Poesio (2005b), to which all distance metrics in (Krippendorff, 2004), and (Geertzen & Bunt, 2010) conform, are adopted here. That is; (1) the distance between a label and itself is 0, and (2), the distance between two labels is not dependent on their order. Because CAMS defines DA and AP, and they combine to form AP-types, it is necessary to define distance functions, such that, the distance of the combined DA and AP label still falls in the range [0, 1] and conforms to the above constraints.

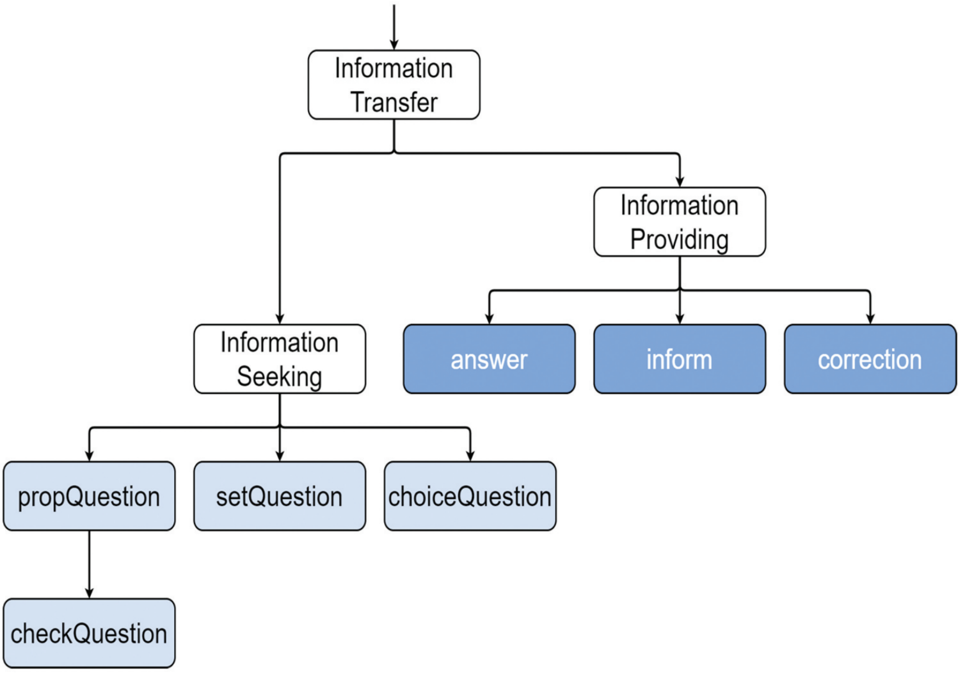


Figure 2. The Information Transfer sub-tree of the DA relationship graph. Leaf nodes are DA, while intermediate nodes represent the communicative function subcategories.

Dialogue act distance function

Geertzen and Bunt (2010), proposed a distance function based on a hierarchical ancestor-offspring relationship between DA labels within the Dynamic Interpretation Theory (DIT^{++}) annotation scheme. Given that DIT^{++} shares many characteristics of the DAMSL scheme (Allen & Core, 1997), and that both of these are precursors to DiAML (British Standards Institution, 2012), a similar approach is employed here. However, their metric considered both the difference in depth and the minimal depth between two labels in the hierarchy, and these are each modified by two constants *a* and *b*. To avoid selecting two arbitrarily chosen constant values, which may affect the coefficient calculation, the DA distance function defined here only considers the distance between two labels within the relationship hierarchy.

The DA relationships are characterized in an undirected graph, where leaf nodes are DA labels and intermediate nodes represent the communicative function subcategories. All edges are considered to have an equal distance of 1. DA are arranged according to their communicative functions which closely match those defined in DiAML. However, in a number of cases DA have been separated into subcategories that more closely resemble their semantic intent. For example, within DiAML the information-providing functions include the DA *agreement* and *disagreement*, which clearly have opposing sentiments, positive and negative. In such cases, DA that are assigned to more appropriate subcategories, for example, positive and negative responses. Figure 2 depicts the Information Transfer sub-tree of the DA relationship graph.[2]

[2] The full DA relationship graph can be found at: github.com/NathanDuran/CAMS-Dialogue-Annotation/blob/master/data_processing/README.md

For each pair of DA, $da_j, da_k \in \mathbf{DA}$, the distance value is calculated as follows. First, the path distance (**p**), between da_j and da_k, is calculated as the sum of the number (**N**) of edges e, each with distance 1, for the shortest path between da_j and da_k:

$$p_{da_j da_k} = \sum_{i=1}^{N} e_i \tag{9}$$

The path distance $p_{da_j da_k}$, is then normalized by the minimum and maximum path distances over the full DA relationship graph, for all possible label pairs ($\mathbf{P}_{min}$ and $\mathbf{P}_{max}$), to yield the distance $d\left(da_j, da_k\right)$, in the range [0, 1]:

$$d\left(da_j, da_k\right) = \frac{p_{da_j da_k} - P_{min}}{P_{max} - P_{min}} \tag{10}$$

Adjacency pair distance function

AP, like DA, can be organized into categories that represent their function: *base, pre, post* and *insert*. However, the paired nature of FPP and SPP, means representing their relationship in a graph-like structure is less appropriate. For example, FPP-pre and FPP-post could be considered similar, in that they both initiate a sequence. Yet functionally, the *pre* and *post* expansion types have opposing meanings, pre-expansions should take place *before* a base pair and post-expansions *after*. Therefore, the distance function defined here considers the difference between the AP labels prefix and suffix, that is, whether they are part of an adjacency pair and initiating or responsive within a sequence (FPP or SPP), or a minimal expansion, and whether they belong to the same *base* sequence or *expansion type* (pre, post and insert).

For each pair of AP, $ap_j, ap_k \in \mathbf{AP}$, the distance value is calculated as follows. First set the distance between ap_j and ap_k to 0, $(d_{ap_j ap_k} = 0)$. Then, separately compare the prefix and suffix of the two labels. If they do not match, increase the distance by .5:

$$d\left(ap_j, ap_k\right) = \sum\nolimits^{0} .5\left(1 - \delta(ap_j^{pre}, ap_k^{pre})\right) + 0.5\left(1 - \delta(ap_j^{post}, ap_k^{post})\right) \tag{11}$$

Thus, two identical AP labels will have a distance of 0, and two completely different labels will have the maximum distance of 1, and two FPP labels will have a distance of .5, as in the previous example with FPP-pre and FPP-post. Similarly, a minimal expansion will have a distance of .5 to the FPP and SPP expansions within the same functional category.

AP-type distance function

Within CAMS, an AP-type label is considered the combination of the DA and AP labels assigned to that utterance, and a similar approach is taken for the AP-type distance calculation. The distance between two AP-type labels is considered the sum of the distances for the individual components, $d\left(da_j, da_k\right) + d\left(ap_j, ap_k\right)$, normalized by the minimum and maximum distances for all possible label pairs ($\mathbf{D}_{min}$ and $\mathbf{D}_{max}$). Thus, for each pair of AP-type labels, $apt_j apt_k \in (DA \cup AP)$, the "raw" distances, $d_{apt_j apt_k}$, are calculated as

$$d_{apt_j apt_k} = d\left(da_j, da_k\right) + d\left(ap_j, ap_k\right) \tag{13}$$

The distance function is then:

$$d\left(apt_j, apt_k\right) = \frac{d_{apt_j apt_k} - D_{min}}{D_{max} - D_{min}} \tag{14}$$

This simple formulation has the advantage of maintaining consistency with the DA and AP distance functions, allowing for comparison of coefficient values between the component label types. Additionally, the large number of possible combinations of DA and AP (297, though not all combinations are valid), would make defining a distinct AP-type distance function laborious and prone to errors and inconsistencies.

Coefficient selection

The following section discusses considerations around the selection of agreement coefficients for calculating inter-annotator agreement. Given that annotators assign DA and AP labels independently, and that each label type has a distinct distance function, it is also possible to calculate independent inter-annotator agreement values for each label type.

The DA within the schema can be grouped into semantically similar communicative functions (Bunt, 2011), such as, information seeking and information providing. Further, some utterances can be thought of as *multidimensional* (Bunt, 2006), that is, they could be assigned two equally valid DA labels (or arguably both). Consider the following example:

```
A1: What is the weather going to be today and tomorrow?
B1: What city would you like to know the weather about?
A2: I want to know if it will drizzle in Durham.
```

Utterance A2 could be considered an answer to the previous question B1, the location they want to know the weather for, or a question in its own right, "*will it drizzle in Durham.*" Clearly, even with well-defined label definitions, there is a certain amount of subjectivity in assigning a single label to certain utterances. A similar semantic grouping is also true for AP, where, for example, FPP-insert and SPP-insert are more closely related to an insert-expansion than AP from the Pre and Post groups. It seems reasonable to treat assignments that belong to different expansion types more seriously than those from the same group. As with DA, there is also an element of subjective interpretation involved when assigning AP labels. For example, identifying which utterances represent the "core action" for a given sub-sequence of dialogue, and therefore should be assigned base-type labels, and those that should be considered expansions. The above, and the use of weighted agreement for DA annotation by (Geertzen & Bunt, 2010), indicates the use of weighted agreement measures, such as Alpha and Beta, are the appropriate choice for DA and AP annotation because the labels are not equally distinct from each other.

What is less clear, however, is the choice between these two coefficients. There has been much debate on this matter (Artstein, 2018; Byrt et al., 1993; Craggs & Wood, 2005; Di Eugenio & Glass, 2004; Hsu & Field, 2003; Krippendorff, 2004; Zwick, 1988). Of course, Krippendorff built the notion of a single distribution into his Alpha coefficient, and Craggs and Wood (2005), argued strongly against the use of coefficients with bias, stating that, "*the purpose of assessing the reliability of coding schemes is not to judge the performance of the small number of individuals participating in the trial, but rather to predict the performance of the schemes in general.*" Yet, Artstein and Poesio (2005b), in their proposal of the Beta statistic believe that, "*assuming that coders act in accordance with the same probability distribution is too strong of an assumption, hence 'biased' measures are more appropriate.*"

The argument against the use of biased coefficients, illustrated by Krippendorff (2004), and others (Byrt et al., 1993; Di Eugenio & Glass, 2004; Zwick, 1988), lies in its calculation of expected agreement. Though biased measures, such as Kappa and Beta, estimate expected agreement on the basis of individual annotator label distributions, they fail to account for unequal distributions *between* annotators. In so doing, biased coefficients effectively discount some of the disagreement resulting from different annotator distributions by incorporating it into expected agreement (Artstein & Poesio, 2008). Thus, for a fixed observed agreement, when annotators produce unequal distributions for the available categories – when bias is present – the values of biased coefficients will *exceed* those of non-biased coefficients. The objection, then, is the "paradox" that as annotators become less similar, biased measures can *increase* (Di Eugenio & Glass, 2004), and begin to diverge from their non-biased counterparts. However, Artstein and Poesio (2005b) point out that in practice the difference between biased and non-biased measures often doesn't amount to much, and that bias is a source of disagreement in its own right. To this latter point, Banerjee et al. (1999), in reference to Zwick (1988), suggested that, "*rather than straightway ignoring marginal disagreement or attempting to correct for it, researchers should be studying it to determine whether it reflects important rater differences or merely random error*." For example, Hsu and Field (2003) demonstrated how Kappa can give useful information even when the individual annotators distributions are very different, and Wiebe et al. (1999), exploited bias to improve the annotation process. In any case, what does seem to be agreed upon, is that as the number of annotators is increased the difference between biased and non-biased measures becomes less significant (Artstein & Poesio, 2005a, 2008; Craggs & Wood, 2005). Further, as stated by Di Eugenio and Glass (2004), the biased and non-biased paradigms reflect distinct conceptualizations of the problem, and in agreement with Artstein and Poesio (2008), the choice should depend on the desired interpretation of chance agreement. However, Di Eugenio and Glass (2004), also believed the bias coefficient (Kappa) is more appropriate for discourse and DA tagging, because "*it is questionable whether the assumption of equal distributions underlying* Pi *is appropriate for coding in discourse and dialogue work*." Yet, they also suggested reporting Kappa and Pi together, to account for the "bias problem" we have just described. Here a similar approach is taken, and both Alpha and Beta will be reported.

Coefficient evaluation

To reiterate, the purpose of measuring agreement for this study is to assess the *reproducibility* of the schema for annotating dialogs with DA, AP and ultimately AP-types. If multiple annotators can be shown to *reliably* assign similar labels to a set of data, it can be inferred that they have a similar understanding of the meaning of the labels, the data items to which they are applicable and that the observed agreement (or disagreement) is not purely a product of chance or an individual's interpretation of the scheme. Unfortunately, the question of what constitutes reliable agreement when interpreting agreement coefficients seems to be an unanswered question (Artstein & Poesio, 2008; Craggs & Wood, 2005; Krippendorff, 2004).

The principal approach is based on a range of values proposed by Landis and Koch (1977). Values below zero are considered "Poor" agreement, and values between 0 and 1 are separated into five ranges: *Slight* (.0 – .2), *Fair* (.21 – .4), *Moderate* (.41 – .6), *Substantial* (.61 – .8), and *Perfect* (>.81). Though they themselves concede that the divisions are arbitrary and only provide a useful benchmark. In Computational Linguistics, it is generally accepted that values of > 0.8 can be considered "good reliability," and values in the range [.67, .8] allow for "tentative conclusions to be drawn" (Carletta, 1996; Krippendorff, 2004). Though it is acknowledged that, as with the original Landis and Koch (1977) values, because of diversity in both the phenomena being annotated and the applications of results, these ranges are not suitable in all cases (Carletta, 1996; Craggs & Wood, 2005; Di Eugenio & Glass, 2004; Krippendorff, 2004). This is especially true for annotation tasks such as this, where there is a degree of subjectivity in choosing an appropriate label, where some prior subject-specific knowledge is required, and notably for AP, prefect agreement will generally require annotators to agree on two (or more) labels, rather than one for DA. Indeed, it has been shown that achieving even the minimum 0.67 value is extremely difficult for discourse annotation (Hearst, 1997; Poesio & Vieira, 1998). This problem is further compounded when using weighted agreement coefficients, because the choice of distance function greatly impacts the calculated coefficient value, as shown by Artstein and Poesio (2005b). Furthermore, regarding the bias problem discussed in the previous section, differences in annotator distributions (bias) will *increase* biased coefficient values, causing them to diverge from non-biased measures. Thus, in the presence of bias, a biased coefficient will always be larger than a non-biased one, and for this reason Geiß (2021) suggests that applying the same range of values is not appropriate, because they warrant different interpretations. Unfortunately, to the best of our knowledge no alternative scale for interpreting biased coefficients has been proposed within the literature, though some have made attempts to "correct" for bias when there are only two categories (Byrt et al., 1993). We therefore choose to evaluate both coefficients, Alpha and Beta, with respect to the ranges typically adopted throughout the literature; with the caveat that, for Beta it is necessary to be cautious when drawing conclusions if there is a significant difference between the two coefficients. Ultimately, choosing an agreement threshold should not be the sole measure upon which an annotation schema, or labeled corpus, should be considered reliable (Artstein & Poesio, 2008; Craggs & Wood, 2005). Instead, the methodology for collecting and calculating reliability should be thoroughly communicated, so that conclusions can be drawn based on the characteristics and motivations of the particular study (Artstein & Poesio, 2008). The following annotation methodology considerations were suggested by Krippendorff (2004, ch. 11), and reiterated by (Artstein, 2018):

(1) Annotators must work independently, so agreements come from a shared understanding not through discussion.
(2) Annotators should come from a well-defined population, so that researchers are aware of previous knowledge or assumptions they bring to the annotation process.
(3) Annotation instructions should be exhaustively formulated, clear and contain step-by-step instructions on how to use it.

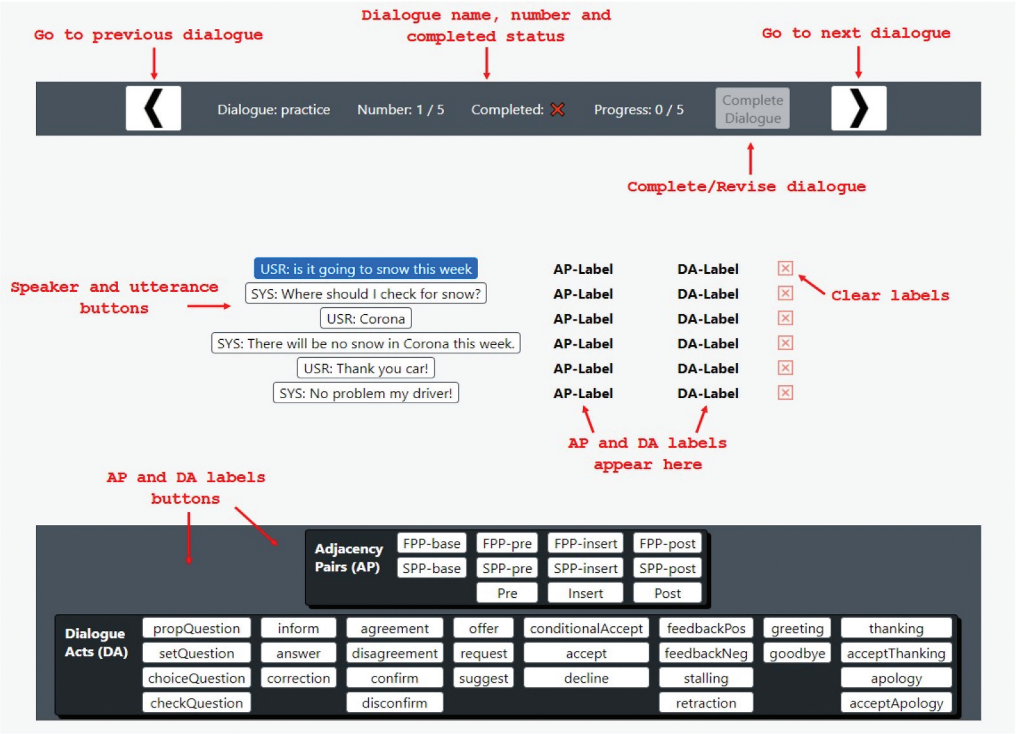
These methodological considerations, and other types of data collected – annotation time and confidence – are discussed in the following section.

Data and methods

The following outlines details of the annotation procedure that was conducted to assess CAMS with respect to; (1) the extent to which multiple annotators agree when applying the schema to dialogue, the inter-annotator agreement, (2), its suitability for application to both task-oriented and non-task-oriented (general talk) dialogs, and (3), evaluate additional characteristics of the material, or annotator behaviors, which may affect application of the schema and the resulting agreement scores. These objectives are intended to establish whether CAMS is comprehensively and explicitly defined, such that it can be reliably applied by multiple annotators, and that it is generalizable to any conversation type, topic, or domain, in order to create corpora annotated with labels that express the syntactic and semantic structure.

The study participants were asked to label five dialogs, containing both task and non-task-oriented conversations, using a specially developed software annotation tool[3] (Figure 3). In total, 15 participants took part in the study (see Participant Selection), and each was assigned one of the five different sets of dialogue for annotation (see Dialogue Selection). The dialogue sets were evenly distributed among the participants, resulting in three annotators per set. The first dialogue in each set is a practice dialogue, followed by the four dialogs in their respective set (two task-oriented and two non-task-oriented). The latter four dialogs were shown to participants in a random order to encourage independent annotation, and mitigate any learning effect of the software, or schema, on annotation results. The participants were given one hour to annotate all dialogs and had no previous training using the annotation tool or CAMS. Upon completion of each dialogue, participants were asked to rate, by means of a Likert Scale, how well their annotations fit the data. Timing data was also collected

[3] The annotation tool, an example of dialogue for each corpus, and all data generated by this study is available at: github.com/NathanDuran/CAMS-Dialogue-Annotation



USR: is it going to snow this week
SYS: Where should I check for snow?
USR: Corona
SYS: There will be no snow in Corona this week.
USR: Thank you car!
SYS: No problem my driver!

Figure 3. Annotation screen of the software annotation tool.

during the annotation process, which recorded how long participants spent annotating each utterance of dialogue. The timing and rating data were used, in addition to the calculated inter-annotator agreement, for further analysis of the manner in which annotators apply the schema, and comparison of task and non-task-oriented dialogs. The following discusses the evaluation measures, and the selection of participants and dialogs in more detail.

Dialogue selection

A key objective of this study is to assess CAMS when it is applied in both task-oriented and non-task-oriented settings. Here, a task-oriented dialogue is defined as, an interaction in which at least one participant has some predetermined goal, such as asking for directions, and engages in the conversation in order to meet that goal. Once that goal is met, or if it is unsuccessful, the interaction is concluded. In contrast, a non-task-oriented dialogue, or general talk, is one in which no participant has a specific predetermined purpose for the interaction other than social communication. Topics may change frequently, and while information may be exchanged it is not in the pursuit of some external predetermined purpose. The dialogs selected for this study are therefore representative of these two groups. Additionally, in order to provide a more representative selection between the groups, dialogs were chosen from four different corpora, with varying numbers of utterances, participants and formats.

In total 20 dialogs were chosen, 5 from each corpus. These were then split into five dialogue sets, each containing one dialogue from each corpus, and grouped in order to keep the total number of utterances in each set roughly equivalent. Additionally, each set contained the same short practice dialogue, selected from the KVRET corpus. The practice dialogue is intended to mitigate any learning

Table 2. Summary of dialogs, and number of utterances, per dialogue set. Total column includes 6 utterances for the practice dialogue.

Set	KVRET	Utts	bAbI	Utts	CABNC	Utts	SCoSE	Utts	Total
1	test 28	7	test 290	7	KB7RE015	9	mammoth	19	48
2	test 52	8	test 428	7	KBKRE03G	6	clone	19	46
3	test 96	4	test 555	5	KDARE00G	4	accident	29	48
4	test 129	6	test 564	5	KE2RE00Y	4	hunter	25	46
5	test 102	4	test 894	5	KBERE00G	5	tipsy	26	46
μ		5.8		5.8		5.6		23.6	46.8

effect associated with the annotation software, and also provide a control dialogue annotated by each participant regardless of the dialogue set they are assigned. Table 2 provides an overview of each dialogue set used within the study. Next is a brief overview of each corpus.

KVRET

Key-Value Retrieval Networks for Task-Oriented Dialogue, is a multi-turn, multi-domain, task-oriented corpus (Eric & Manning, 2017). The data was collected using a Wizard-of-Oz scheme, via 241 workers on Amazon Mechanical Turk. It contains 3,031 dialogs in 3 domains for an in-car personal assistant: calendar scheduling, weather information and point-of-interest navigation. The dialogs used for this study were randomly selected from the 304 dialogs in the KVRET test set.

bAbI

The Dialogue bAbI Tasks data is a subset of the bAbI project by the Facebook AI Research group (Weston et al., 2015). The set of six tasks are designed to test end-to-end dialogue systems in the restaurant booking domain (Bordes et al., 2017). The dialogs used for this study were randomly selected from the 100 dialogs in the bAbI task 1 test set. Each dialogue follows a similar format. First greetings are exchanged, and the automated system asks the user what it can help them with. The user states their preference of cuisine, location, price range, and number of diners, and in some cases extra system turns clarify these preferences.

CABNC

The Jeffersonian Transcription of the Spoken British National Corpus is a conversation analytic re-transcription of naturalistic conversations from a sub corpus of the British National Corpus (Albert et al., 2015). It contains 1436 conversations with a total of 4.2 million words. There is a wide range in the number of utterances within the CABNC dialogs, in many cases hundreds or thousands of utterances. In order to, as much as possible, maintain a similar number of utterances across all dialogs and dialogue sets, and due to time constraints, those used for this study were randomly selected from dialogs with less than 10 utterances.

SCoSE

The Saarbrucken Corpus of Spoken English consists of 14 transcribed dialogs of general talk on a range of topics between two or more participants (Norrick, 2004). As with the CABNC corpus, due to the large number of utterances, and time constraints, those chosen for this study were the 5 dialogs with the fewest utterances. In our set, the *mammoth, clone*, and *accident* dialogs take place between up to three undergraduate students sharing an apartment, while *hunter*, and *tipsy* take place between Helen and her three adult daughters before a late-afternoon Thanksgiving dinner.

Participant selection

The study participants comprised of 15 undergraduate students from the 1st year of an English Language and Linguistics course. For 5 weeks prior to the study the participants received instruction on CA and AP as part of their linguistics syllabus. However, we also wanted to assess how intuitive the schema is to apply with only minimal prior knowledge. Given its purpose is for computational dialogue modeling, CAMS should ideally be usable by as wide a range of people as possible. Not only Conversation Analysts, but Computer Scientists, Computational Linguists, and other NLP practitioners, who either already have some familiarity with CA and AP, or who simply intend to follow the annotation guidelines and label definitions. This is particularly important when considering the application of the schema for further annotation tasks, for example, creating large datasets for training and evaluating deep-learning NLP models. Therefore, our participants were not provided any specific instruction regarding CAMS and did not receive any training in its application. As such, participants could reasonably be considered novice annotators, in that, they had some prior knowledge of CA theory but no previous experience in annotation or applying CAMS. The selection of Linguistics students as annotators was largely for pragmatic reasons:

(1) While DA labels could be considered somewhat intuitive, even for novice annotators, AP require some level of previous CA knowledge. Therefore, conducting a large-scale crowed-sourced annotation experiment, where we cannot guarantee any prior understanding of CA concepts, would be inappropriate.
(2) Even though expert annotators are more likely to produce high agreement (Geertzen et al., 2008; Nowak & Rüger, 2010; Snow et al., 2008), the number of available expert annotators is limited. Further, both Krippendorff (2004), and Carletta (1996), argue that, for discourse and dialogue annotation schemes there are no real experts, and that what counts is how totally naïve annotators manage based on written instructions. While using naïve annotators is not appropriate here, the use of non-expert annotators should still provide some insight into the clarity of the CAMS label definitions and annotation guidelines.
(3) Bayerl and Paul (2011), suggest using annotators with the same level of domain expertise. Using participants from the same student cohort, with a similar level of experience, should therefore reduce external factors which may influence the interpretation of the schema definitions and guidelines.

Timing and rating measures

The annotation tool collected additional utterance annotation timing and label confidence data for each annotator. The purpose is to augment the comparison between task-oriented and non-task-oriented dialogs, and the different label types within the schema, that would not be possible with agreement coefficient data alone. It also provides additional insight into the participants annotation behavior, such as a change in confidence, or the amount of time spent selecting labels, which may indicate how well annotators are able to learn and internalize the annotation scheme.

Annotation timing

The annotation software allows users to select an utterance of dialogue, which is then highlighted to signal it is the “target” for annotation. With an utterance selected, the user chooses a single DA and AP label to assign by clicking on their respective buttons. An utterance is considered *labeled* when it has been assigned one of each label type. At which point the software automatically selects the next unlabeled, or partially labeled, utterance. The time taken to annotate an utterance is measured as the total time the utterance is selected and unlabeled. This time is cumulative, so if a previously assigned label is removed, so that a different label can be selected, or it is unselected and re-selected later, any further annotation time is added to the previous total.

Annotation confidence

Once a dialogue is fully labeled users are presented with a questionnaire screen. Here, they are asked to rate how well their assigned labels fit the dialogue in question. Ratings are provided by means of a Likert Scale between 1 and 7, with 1 representing *not at all*, and 7 *perfectly*. There are three questions, one for each label type; and the prompts emphasize the purpose of these label types. For example, how well the DA describe the communicative *meaning* of the utterances, AP the *structure*, and for AP-types, how well they combine to convey both structure and meaning. In addition to the confidence ratings, users are given the option to highlight any of the labels they assigned to the current dialogue. This is because users must fully label each utterance, there is no option to leave an utterance unlabeled, or partially labeled, and therefore provides an opportunity to indicate whether they feel certain labels did not adequately described the utterance, or selection of utterances.

Statistical analysis

Throughout our analysis we perform hypothesis testing in the form of Two-sided t-tests or Analysis of Variance (ANOVA), where appropriate. Where the results of an ANOVA reveal a significant overall effect, we perform a further Tukey's Honest Significant Difference (Tukey-HSD) post-hoc analysis, in order to determine the factors contributing to the observed effect. Due to relatively small sample sizes, we calculate the ω^2 effect size and adopt the standard ranges for interpretation, low (.01 – .059), medium (.06 – .139) and large (.14+). For t-tests we report Cohen's d effect size, with standard interpretations of small (.2), medium (.5), and large (.8+). Throughout the analysis, we use a significance level $\alpha = .05$, and, unless otherwise stated, the statistical power is $\geq .8$.

Results and discussion

In this section the results of the annotation procedure are presented and some of the observations that arise are discussed. We begin with the inter-annotator agreement measures, firstly for each set of dialogue, before examining agreement for task and non-task-oriented dialogs, and each corpus. We then report the results for annotator confidence and timing data, respectively.

Inter-annotator agreement

Inter-annotator agreement was calculated for the Alpha and Beta coefficients from the recorded annotations for each dialogue set. Figure 4 shows agreement values for each label type (DA, AP, and AP-type), and the overall mean agreement for each coefficient.

Figure 4 and subsequent statistical analysis show that:

- According to the Landis and Koch (1977), scale we find that agreement for the Beta metric is "substantial" for DA (.74) and AP-types (.67), and "moderate" (.6) for AP alone. Using the range [.67, .8] (Carletta, 1996; Krippendorff, 2004), we find that only DA and AP-type labels are able to reach this threshold for the Beta coefficient.
- The Alpha metric produces the same pattern, but with lower values of agreement. DA agreement is 'moderate' (.47), while AP are 'slight' (.18), and AP-types 'fair' (.33). Comparing Alpha and Beta values, for each label type, show these are all significantly different ($p < .001, d > 1$). Possible reasons for this are explored further in section Alpha vs Beta.
- ANOVA over the label types (DA, AP, and AP-type) for each metric showed large effect sizes ($\omega^2 = .186$ and $\omega^2 = .179$ for Alpha and Beta respectively). Post-hoc analysis, reveals that this arose almost wholly from the AP:DA difference ($p < .001$) for both metrics.

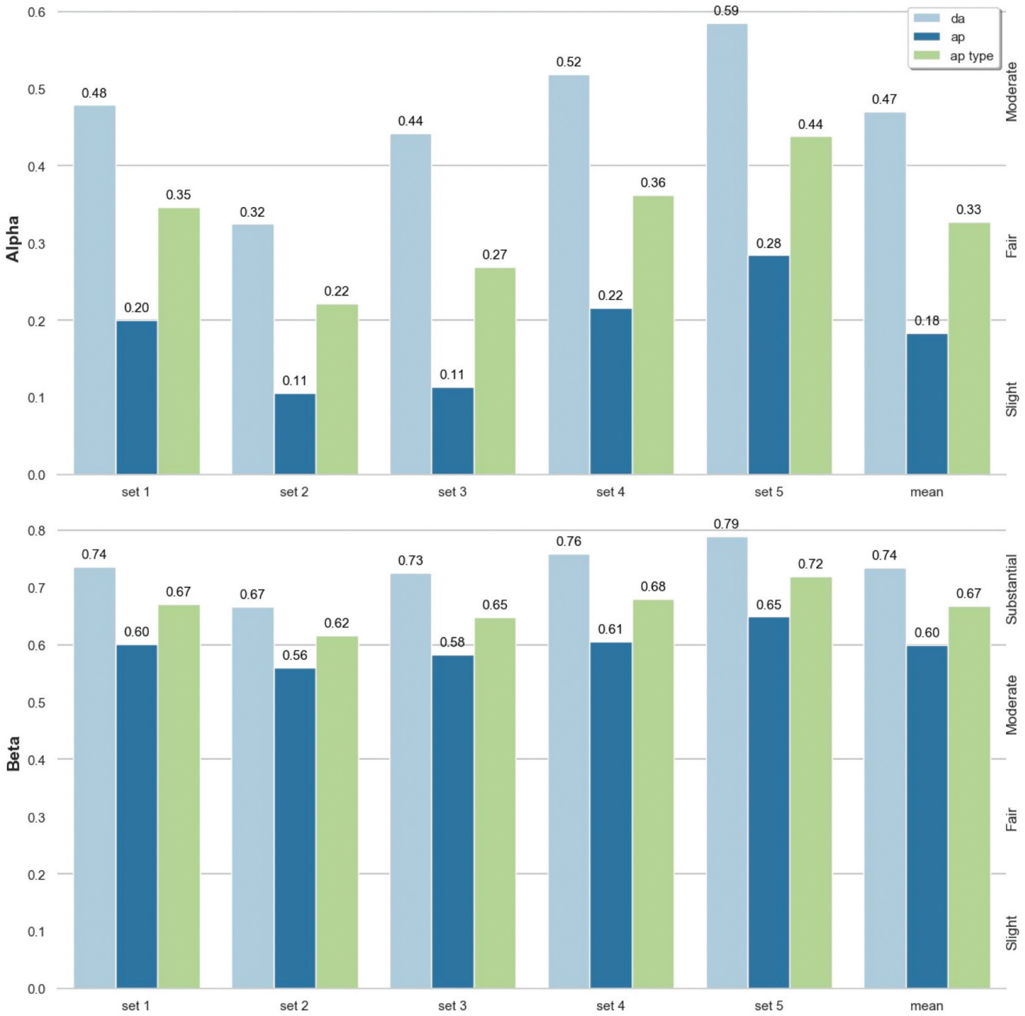


Figure 4. Alpha and Beta inter-annotator agreement values for each dialogue set.

Overall, we see a considerable difference between the values of Alpha and Beta. Though it is less pronounced for DA labels, with a mean difference of 0.27, than it is for AP, and AP-types, which differ by 0.42 and 0.34, respectively. These differences indicate that annotators had very different proclivities when assigning labels, and this bias has *increased* the values of Beta with respect to Alpha. In the case of AP this increase amounts to two full thresholds on the Landis and Koch (1977) scale, from “slight” to “moderate,” and we therefore recommend that this is considered before drawing any conclusions of reliability from the Beta agreement values alone. However, that this difference is less for DA, and greater for AP, suggest that individual annotator distributions were more similar when assigning DA labels and less similar for AP labels. In other words, we see a higher degree of idiosyncratic interpretation between the annotators when selecting AP labels, and this is reflected in the difference between the two coefficients. This observation is discussed further in AP Label Agreement and Alpha vs Beta.

Task-oriented and Non-task-oriented dialogs

A primary focus of this study is to investigate the extent to which the schema can be applied to different types of dialogue. Annotated dialogs were therefore split into their respective task and non-task-oriented groups, and again agreement was calculated using Alpha and Beta for each label type. Figure 5 shows the resulting agreement values for each dialogue group, and the practice dialogue:

- On the practice dialogue, the Beta metric reports "perfect" agreement for all three groups of labels on the Landis and Koch (1977), scale (Beta > .95).
- For the Alpha metric, agreement on the practice dialogue is again "perfect" for DA (.84), and high for the AP-types (.59) but lower for just the AP labels (.37).
- These practice results are consistently higher than the main results, possibly because there are more annotators, and (as will be seen later) due to the nature of the KVRET corpus.
- Agreement was consistently higher for task-oriented dialogs for all label types, and both coefficients. Overall, these differences are statistically significant ($p < .001, d > 1$) for both Alpha and Beta. Only when looking at just the AP labels, is the task vs. non-task distinction not statistically significant ($p = .07, d = .86$ and $p = .56, d = .9$ for Alpha and Beta, respectively).

Again, overall, the differences between the two coefficients is high in most cases, and consequently we advise caution when interpreting the Beta values with respect to typical agreement thresholds. However, it is worth noting that for DA labels the difference on the task-oriented dialogs (0.19), and the practice dialogue (0.15), is much smaller than previously observed. Therefore, we can conclude that, not only is agreement higher but individual annotator distributions were more similar.

To examine the difference between the task-oriented and non-task-oriented groups further, Table 3 shows the assignments produced by two annotators, users 10 and 5, for a task (KVRET) and non-task (CABNC) dialogue. We selected users 10 and 5 for this analysis because both exhibit a competent understanding of CAMS and its application. Yet as we will see, their differing interpretations of the CABNC dialogue led to negative agreement values. On the other hand, for the KVRET dialogue they reached near perfect agreement. Thus, this pairing provides clear insight into the properties of task-oriented and non-task-oriented dialogs that contribute to the observed differences in agreement between these groups, even between annotators who demonstrate a similar understanding of the annotation scheme. Additionally, both annotators made some small errors in assigning AP or DA. We highlight these assignments here and explore some of these observations further in the AP Label Agreement section.

Firstly, we can see both annotators assign an invalid AP label to utterance A3; user-5 begins a FPP-post without a closing SPP, and user-10 places an insert label *outside* of a FPP/SPP base-pair. User-10 also incorrectly begins a FPP-pre (A1) and closes with a SPP-base (B2), a pattern that is repeated in the KVRET dialogue. There are also some minor misuses of DA. In particular, user-5 assigns "stalling" to (A2), which represents a speakers need for a little extra time to construct their contribution, for example, "Let me see. . . " or "Umm. . .." Given the nature of the following utterances, a question-type DA, or user-10's assignment of negative feedback, is more appropriate. However, the assignment of negative feedback for A3 is certainly incorrect, as this DA represents the speakers mishearing, or misunderstanding, of the previous utterance; a conclusion that is not borne out by its content.

Regarding AP, the main source of disagreement with the CABNC dialogue is what constitutes the core action or communicative goal, and thus should be assigned as base-type AP, and what utterances contribute to, or support, this action, and should therefore be expansions. Both correctly identify the core action as a request to turn the radio off in A1. However, user-5 considers this action complete with the refusal to do so in B1, and the following two utterances are merely clarifying the meaning of "whatsname." On the other hand, user-10 considers that the response in B1 was a mishearing, or misunderstanding, by A and that this requires the insert-pair before the action is completed in B2. Clearly these two interpretations led to significant disagreement between the two annotators and is largely driven by the ambiguity of certain utterances within the transcription, particularly A2. If A2

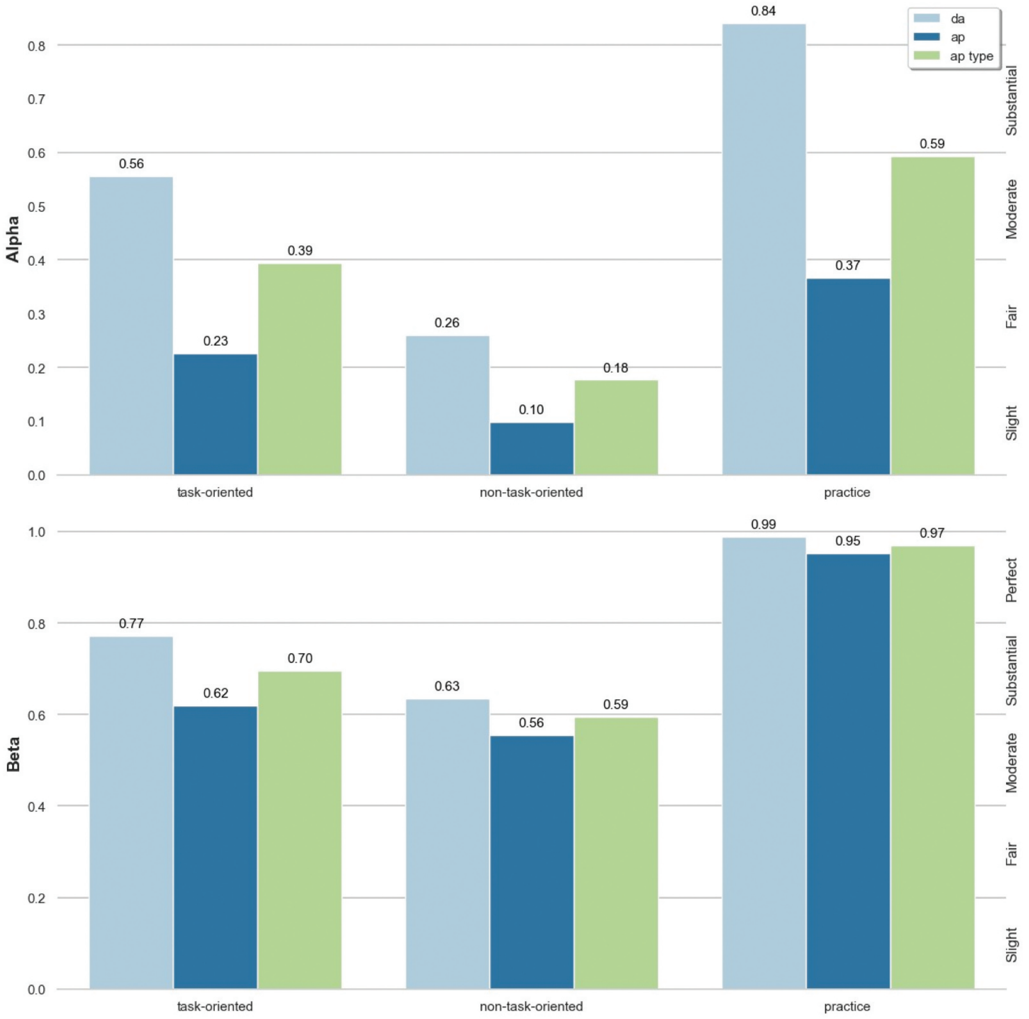


Figure 5. Alpha and Beta Agreement values for task and non-task dialogs.

Table 3. Label assignments by users 5 and 10 for a task (KVRET) and non-task (CABNC) dialogue.

CABNC (KBERE00G)	User-5	User-10
A1: Can you turn that radio off I want to listen to the phone in.	FPP-base propQuestion	FPP-pre request
B1: I got the whatsname on.	SPP-base decline	FPP-insert inform
A2: What What.	FPP-post stalling	SPP-insert feedbackNeg
B2: The whatsname Don't ask me I du n no what it's called.	SPP-post confirm	SPP-base answer
A3: What do you want that on for I'm trying to listen to the radio I want to listen to the phone in.	FPP-post feedbackNeg	Insert disagreement
KVRET (Test 102)	User-5	User-10
C1: Can you find out the date and parties attending my dinner?	FPP-base setQuestion	FPP-pre propQuestion
D1: Your dinner is on Tuesday with your sister.	SPP-base answer	SPP-base inform
C2: Thanks.	FPP-post thanking	FPP-post thanking
D2: you're welcome	SPP-post acceptThanking	SPP-post acceptThanking

were instead “the what?”, or “who?”, then user-5’s interpretation is preferred, or alternatively, “sorry what?”, might suggest user-10’s understanding was correct. Unfortunately, “what what” lends itself to both these possibilities and hence the alternative interpretations. This is also reflected in the negative agreement scores between these two annotators, with an Alpha of −.1, and a Beta of −.05. For the KVRET dialogue there is no such ambiguity in which utterances make up the core action, and this resulted in “perfect”, or near perfect, agreement of .8 and .77 for Alpha and Beta, respectively.

For DA, we again see considerable disagreement for the CABNC dialogue, and this is largely driven by the alternative interpretations previously discussed. Of note, however, is the assignments of a “propositional question” and a “request” for utterance A1. Even though it is posed as a question, this statement is an indirect way of requesting that the radio be turned off, and therefore user-10’s assignment is more suitable (Bunt, 2017). Yet, it is easy to see how a propositional question, which suggests a positive (accept) or negative (decline) answer, is a reasonable alternative interpretation. Interestingly, despite the similar form of utterances A1 and C1, neither annotator assigned the same DA label. These dialogs were not presented in the order shown here, but this does indicate a change, or inconsistency, in interpretation; perhaps influenced by the presence of an interrogation mark in C1 which implies a question-type DA is appropriate. For the CABNC dialogue we again see negative agreement, −.03 and −.06, and for the KVRET dialogue substantial agreement of .79 and .76 for Alpha and Beta, respectively.

From these results, we can see that, while there is some incorrect usage of both AP and DA, the main source of disagreement stems from difficulties interpreting the non-task-oriented dialogue. The two alternative views discussed above suggest two different sets of AP assignments, depending on where one considers the core action to have been completed, and this is largely driven by the ambiguity of utterance A2 observed above. Macagno and Bigi (2018), referred to this phenomenon as “imaginary ambiguity”, that is, a particular utterance can have multiple distinct interpretations for the intended effect on the recipient depending on the context. In this case, A2 is interpreted differently depending on the reading of B1 as a refusal, or misunderstanding. This kind of *meaning multiplicity* (Boxman-Shabtai, 2020) may arise, at least in part, from the nature of transcribed material of natural conversations, where social cues, such as prosody, intonation, and body language, are lost. Indeed, Collins et al. (2019), were able to show that disfluencies in speech can have very different meanings when presented in spoken and written form, and we surmise that this is also true of illocutionary ambiguous utterances. As noted by Green et al. (1997), “*a transcript is a text that ‘re’-presents an event; not the event itself*”, thus information is inevitably lost. In any case, these differing interpretations are a clear example of bias on the part of individual annotators, and have therefore contributed to the inflation of the Beta coefficient, and its divergence from Alpha, that we have previously discussed. On the other hand, for the task-oriented dialogue there is a clear delineation between the core action and the remaining “thanking” utterances. This concurs with the work of Grosz (2018), who established that task-oriented dialogs are structured, with multiple utterances grouping into a dialogue segment, and their structure mirrors the structure of the task. This characteristic simplifies the identification of AP and we therefore see much higher agreement and lower bias.

Corpora dialogs

An additional factor which may contribute to the observed difference in agreement between the task and non-task dialogue groups is the number of utterances in each dialogue. Dialogs in the SCoSE corpus contain an average of 23.6 utterances, around half of the total number of utterances in each dialogue set, and may therefore be contributing a disproportionate amount of agreement (or disagreement) to the overall agreement values. Hence Figure 6 breaks the comparison into different corpora. A further ANOVA and post-hoc analysis of agreement between pairs of corpora, was performed for each label type and coefficient:

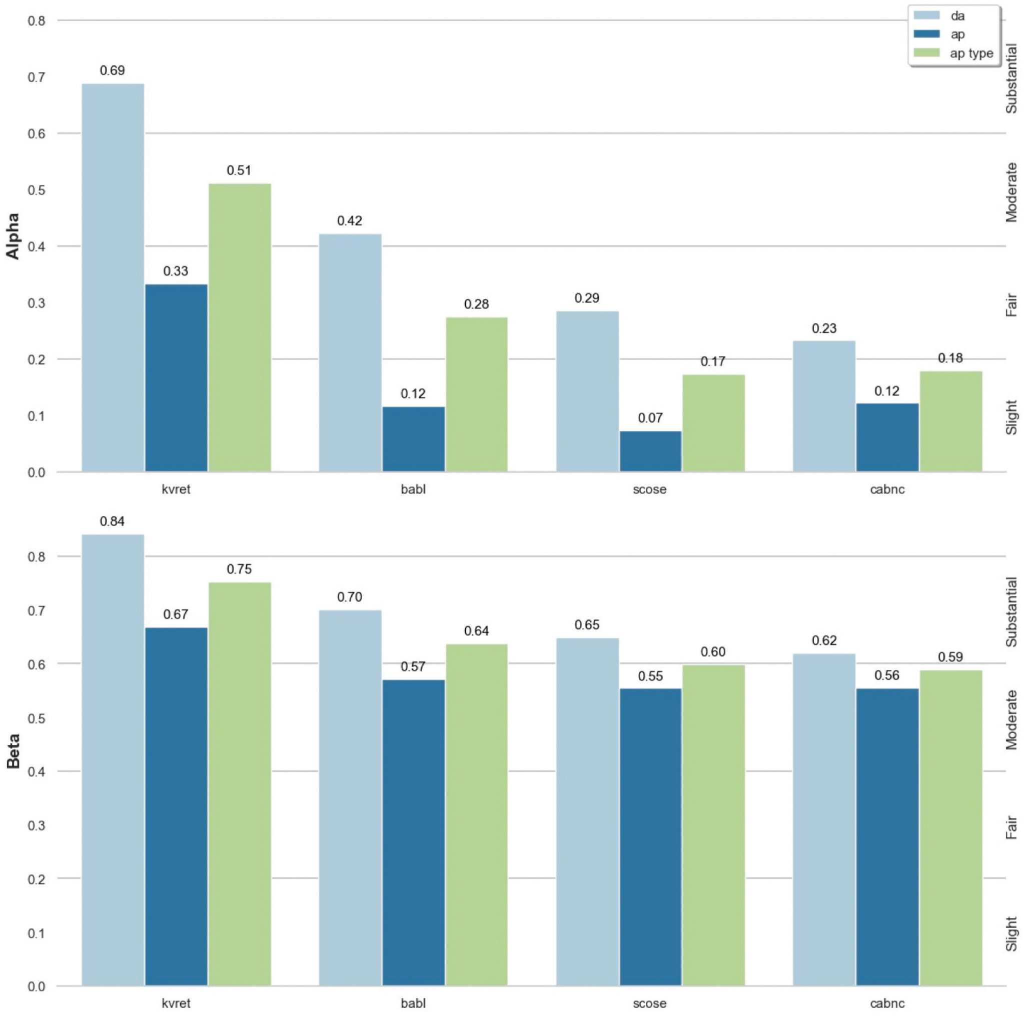


Figure 6. Alpha and Beta Agreement values for each corpus.

- The post-hoc analysis reveals that there is no significant difference in agreements (p = .9) between the two non-task-oriented corpora, CABNC and SCoSE, for both Alpha and Beta coefficients, despite a mean utterance length of 5.6 and 23.6, respectively. This is also the case when comparing the bAbI corpus (mean utterance length 5.8) and the non-task-oriented corpora. Therefore, it is unlikely that the number of utterances is contributing to the observed differences in agreement between the groups.
- Predominantly, the statistically significant results are for DA and AP-type labels between KVRET and the other corpora. This indicates that the difference in agreement values are a product of higher agreement for the KVRET corpus, rather than a difference between the groups. Certainly, agreement is higher on the KVRET corpus, for all label types and both agreement coefficients.
- These results also provide some insight into the previous observation, that there is no significant difference in agreement for AP labels between the groups. Only the KVRET and SCoSE comparison for the Alpha metric produced a significant result (p = .028) and in all other cases we still see no statistical difference for AP labels.

From these results, we can see that, once more, there is a large difference between Alpha and Beta, and this is greater for AP than DA, hence a larger degree of idiosyncratic interpretation between the annotators. However, in accordance with the previous remarks, this bias is lower for the KVRET corpus than it is for the other three. Thus, while agreement for DA is higher for both task-oriented corpora, for AP we see no difference in agreement between the bAbI corpus and the two non-task-oriented corpora.

Dialogs in the bAbI corpus all follow the same basic format. First greetings are exchanged, and the automated system asks the user what it can help them with. The user states their preference of cuisine, location, price range, and number of diners. The system then either asks for clarification of one of the stated preferences, or confirms the preferences are understood, and finally states that it will "look into some options" for the user. As an example, the following is the bAbI test 894 dialogue:

```
A1: good morning
B1: hello what can i help you with today
A2: may i have a table in a cheap price range in london with spanish food for two
B2: i'm on it
B3: ok let me look into some options for you
```

Given that this structure is common to all bAbI dialogs we were able to examine the assignments across all participants and identified common sources of disagreement. For AP, the main source of disagreement is which utterances constitute the core action or communicative function of the dialogue. With bAbI, we see two common interpretations. Six of our annotators considered the core action to begin with utterance B1 and the systems' question of "what can i help you with today", thus assigning B1 and A2 as a *base-type* AP. The remaining annotators all considered B1 as part of the preliminary salutations and assigned a *pre-type* AP label to B1. This latter group therefore began the base-pair from A2 and concluded it at B2 or B3. It is easy to see how these two interpretations can be reached given the multidimensional nature of utterance B1 (Bunt, 2006), that is, both a greeting *and* a question. Though, only two annotators assigned a *greeting* DA label to B1 and eleven assigned a *question-type* label (the remaining two incorrectly assigned an offer label). The multidimensional nature also extends into the interpretation of AP. The greeting component of B1 is responsive to the greeting in A1, indicating it is the *concluding* utterance of a pair, while the question component creates the expectation of a response, suggesting it is *initiating* an AP. Hence, we see two valid readings of the utterances DA, and its relationship to the surrounding utterances, which is reflected in two different interpretations of the core action underway. Therefore, we can see that just as the semantically ambiguous utterance discussed in the previous section led to two valid interpretations of the dialogue, here a similar effect is caused by the multidimensionality of B1, resulting in a significant number of disagreements for AP on the bAbI corpus. Additionally, the multidimensional nature of utterances like B1 are likely to be a further contributor to the bias, and inflation of the Beta coefficient, that we have observed throughout our results.

AP label agreement

As previously observed, there appears to be no significant difference in agreement for AP labels between the task and non-task dialogue groups, and further, that much of this is caused by the negligible difference between the bAbI, CABNC and SCoSE corpora. Manual inspection of the annotations revealed that a considerable amount of confusion seemed to arise around the valid use of FPP and SPP for AP. Often annotators would assign a SPP to initiate a sequence (rather than a FPP), or fail to create a valid sequence entirely, for example, by assigning a FPP without an accompanying SPP. This observation was explored further using an adjusted AP distance function, which ignores the AP prefix (FPP/SPP), and instead *only* considers the difference between the AP *base* or *expansion types* (pre, post, and insert). The "suffix-only" distance function treats all labels as equally distinct, with a distance of 1 for non-identical labels, and 0

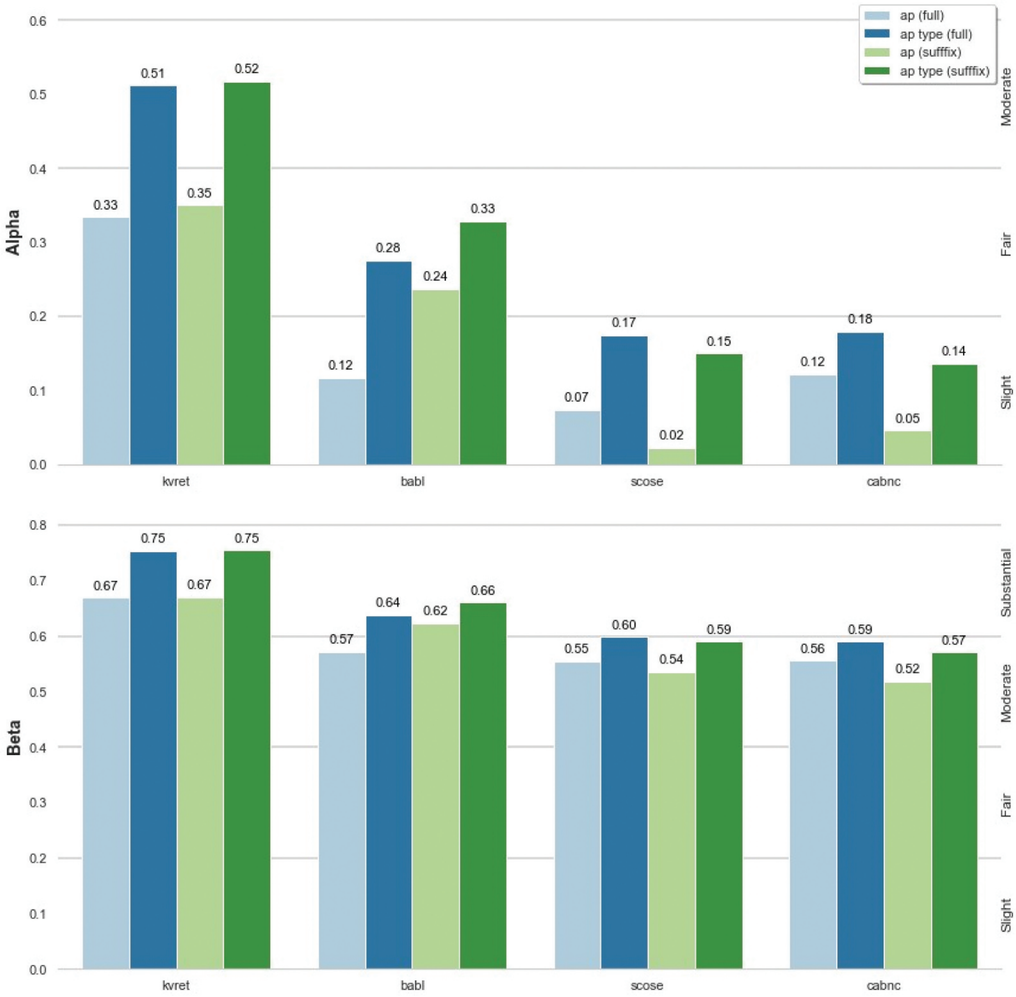


Figure 7. Corpora agreement values calculated with the suffix-only AP distance function.

otherwise. For example, two *insert* type labels (FPP-insert, SPP-insert or insert) would have a distance of 0 between them, but a distance of 1 with all other AP label types. Therefore, the suffix-only distance function should indicate the extent to which annotators misunderstanding of the valid use of FPP and SPP labels contributed to the observed AP agreement values. Figure 7 shows the agreement values that were recalculated for using the suffix-only distance function.

- Using the suffix-only distance function both task-oriented corpora show improved agreement for AP labels, with a minimal improvement for the KVRET corpus but a considerable improvement for bAbI. For Alpha the bAbI agreement doubled from .12 to .24, and Beta shows an increase of .57 to .62.
- Both non-task-oriented corpora show a decrease in AP agreement, though, again the effect is greater for the Alpha coefficient, with a decrease of .05 and .07 for SCoSE and CABNC respectively, compared to .01 and .04 for Beta.
- Post-hoc analysis reveals there is now no longer a significant difference in AP-type labels when comparing the KVRET and bAbI corpora (p = .181 and p = .193, for Alpha and Beta, respectively).

This indicates that, when annotators misunderstanding of the valid use of FPP and SPP is not considered, they tend to more often agree on the *base* and *expansion types* of AP labels for task-oriented dialogs. Whereas, for non-task-oriented dialogs the opposite is true, with a *decrease* in agreement that suggests annotators rarely agree on the AP *base* or *expansion types*. Perhaps unsurprisingly, this suggests that the structure of non-task-oriented dialogs is less well defined, and open to more subjective interpretation, than that of task-oriented dialogs. It may also offer explanation for the lack of significant difference in AP agreement, and high bias, that was previously observed. Using a two-sided t-test to compare the suffix-only agreement scores for AP labels between the task and non-task groups now results in a statistically significant difference for Alpha and Beta ($p = .0028, d > 1$ and $p = .0089, d > 1$, respectively). Therefore, the incorrect usage of FPP and SPP was *reducing* agreement for task-oriented dialogs, while for non-task dialogs *increasing* agreement, and "evening out" AP agreement values between the groups. These results also suggest that using non-expert annotators may not be suitable for this task, as many seem to lack a clear understanding of the proper use of AP, or alternatively, more training beforehand may help to improve understanding in this regard. It is also possible that some of the confusion was caused by the similarity between FPP and SPP, with only one-character difference between the two labels. Perhaps changing the labels to, for example, "first-part" and "second-part," would help mitigate the problem of assigning these in the wrong order.

Alpha vs beta

Previous results have shown that in all cases the Beta coefficient results in significantly higher agreement values than Alpha, and that this is principally caused by the differences in annotator label distributions increasing the Beta values. As discussed in the Inter-Annotator Agreement section, the difference between these two coefficients lies only in their calculation of expected disagreement. That is, Alpha estimates disagreement on the basis that all annotators assign labels with the same probability distribution, while Beta considers the individual annotators distributions. Here, these different estimations are tested, using the actual annotator label distributions from this study, to determine the extent to which annotators use similar, or different distributions.

Jensen-Shannon divergence. The difference, or similarity, between probability distributions can be calculated using the Jensen-Shannon divergence (JSD) method. Here, the generalization of JSD is adopted, which calculates a distance value between two or more probability distributions. The distance value is bounded in the range $0 \leq JSD \leq log_2(n)$, where n is the number of input distributions; the lower bound represents identical distributions and the upper bound maximally different distributions. For each dialogue set the JSD distance was calculated for the probability distributions of all annotators that labeled that set. Thus, in each case $n = 3$ and the range is $0 \leq JSD \leq 1.58$. Table 4 shows the JSD distances for the DA and AP label distributions over each dialogue set. We can see that both DA and AP have low distance values, within $\sim \frac{1}{6}^{th}$ of the lower range, and therefore, overall differences between annotator distributions is relatively small using this measure. AP labels show a lower average distance

Table 4. JSD distance for DA and AP labels of each dialogue set.

Group	DA	AP
set 1	0.272	0.15
set 2	0.305	0.177
set 3	0.183	0.307
set 4	0.232	0.17
set 5	0.26	0.296
μ	0.251	0.22
σ	0.041	0.067

than DA over all dialogue sets, with a mean of 0.22 and 0.25 respectively, which is likely due to the fewer number of AP labels. However, AP also show a higher standard deviation than DA and this may reflect the higher disagreement and bias for AP labels that was previously observed.

Pearson's Chi-squared. In addition to calculating the distance between groups of annotator probability distributions, we can also examine the extent to which label distributions are dependent on the individual annotators that assigned them. For this purpose, an χ^2 test was conducted using the *cumulative* annotator label distributions. For each dialogue set a separate χ^2 test was performed for all pairwise annotator combinations.[4] From these results, we can see that

(1) For DA, in none of the pairwise comparison between annotators are the observed label frequencies significantly different. In other words, regardless of which annotator assigned the labels, the distribution would still be largely the same – although individual assignments could still be very different.

(2) For AP, in $\frac{1}{3}$ of cases (2 in set 3 and all of set 5), we see significant results when comparing the critical value to the test statistic, and also significant p-values. As such, we must reject the null hypothesis and concluded that the label distributions (in $\frac{1}{3}$ of cases) were dependent on the annotator that assigned them. Therefore, certain annotators were producing label distributions that were quite distinct from each other.

These two conclusions seem to support the results from the JSD comparison. Firstly, there seems to be less variance in the annotator's DA label assignments, likely contributing to the observed higher agreement values. Secondly, AP seem to be more dependent on the individual annotator which assigned them (overall p-values are lower, indicating a higher degree of idiosyncratic interpretation). As such, agreement for AP was lower, while bias was higher, and this may also be indicative of the misunderstanding surrounding the use of FPP and SPP that was discussed in the AP Label Agreement section, and the differences in interpretation observed in the task-oriented and non-task-oriented and corpora results. These results also suggest that both the JSD and χ^2 tests could serve as additional measures for the homogeneity of annotators interpretation, and understanding, of the material and coding scheme.

From these measures, and regarding Alpha and Beta, it seems that annotators do, in fact, use more similar distributions for DA labels. In most cases, this also appears true for AP, though there is a greater variance (in part due to misunderstanding FPP and SPP) between some groups of annotators. However, as we have seen, these small differences can result in drastically different values between the two coefficients. Given that there is a certain amount of semantic interpretation when assigning both DA and AP labels, the assumption that annotators will use the same distribution is, as Artstein and Poesio (2005b) stated, too strong. Consequently, Alpha may be too harsh in its estimation of annotator distributions and punish individual interpretation too severely. Yet, as shown in our AP label agreement results, when using the suffix-only distance function, the Beta coefficient exhibited smaller changes in agreement values. Further, as shown throughout our results, in the presence of bias – which is itself a form of disagreement – the Beta coefficient is consistently higher than Alpha. Therefore, it may be a less sensitive measure of agreement, even hiding some causes of disagreement, which makes drawing conclusions of reliability problematic, using the Beta coefficient alone. However, that Alpha and Beta diverge, and the extent to which they do, can provide useful information in its own right. In our case it has clearly signified the higher degree of idiosyncratic interpretation between annotators when assigning AP labels, and also highlighted differences between task and non-task-oriented, or dialogue corpora, groups. This information would not have been apparent from the calculation of either coefficient alone, and so in agreement with Di Eugenio and Glass (2004), for annotation that require a high degree of semantic interpretation, it seems more helpful to report both biased and unbiased values. Though, if the goal is to reach high agreement values, and hence reliability of labeled data, the more stringent unbiased coefficient should be used.

[4] Chi-squared results table is available in full at: github.com/NathanDuran/CAMS-Dialogue-Annotation/blob/master/data_processing/results.ipynb

Table 5. Mean and standard deviation of confidence scores by label type, corpus, and dialogue type.

	KVRET		bAbI		Task		SCoSE		CABNC		Non-task	
Label	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
DA	5.06	1.03	4.53	0.99	4.8	1.03	4	1.31	4.13	0.64	4.07	1.01
AP	5.27	1.09	4.13	0.99	4.7	1.18	3.93	1.16	4	0.85	3.97	0.99
AP-Type	4.87	0.99	4.53	0.96	4.7	0.95	3.67	0.98	3.8	0.68	3.73	0.83
Overall	5.07	1.03	4.4	0.96	4.73	1.05	3.87	1.14	3.98	0.72	3.92	0.95

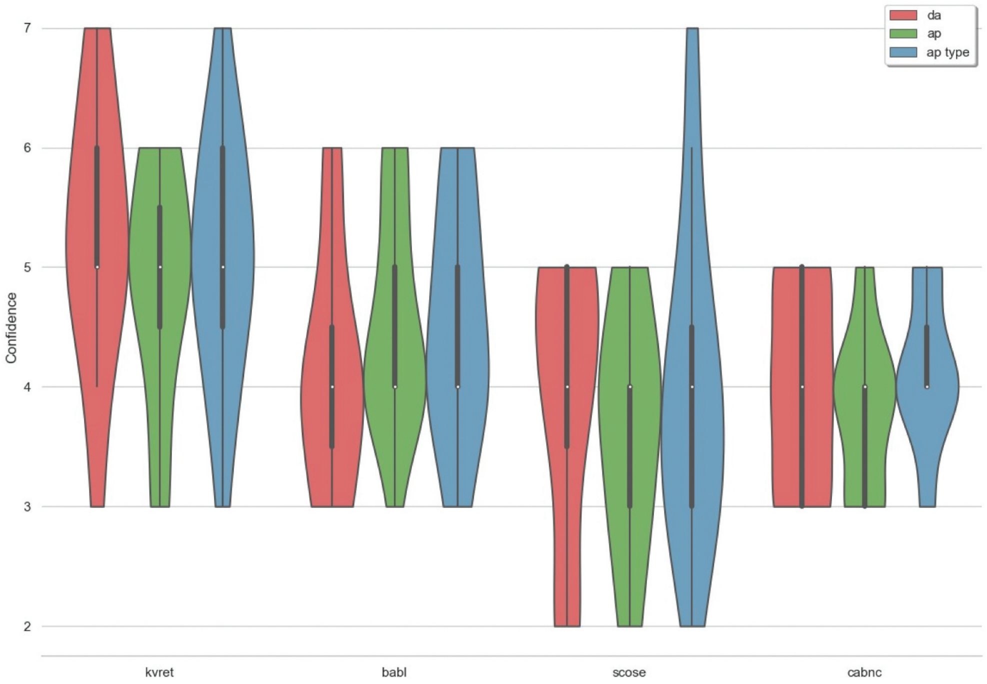


Figure 8. Reported annotator confidence scores for each dialogue and label type.

Annotation confidence scores

Analysis of participants confidence scores supports some of the observations from the previous sections. Overall, annotators reported a higher confidence in their assigned labels for task-oriented dialogs than for non-task-oriented dialogs (Table 5), which coincides with the higher agreement for task-oriented dialogs observed in our previous results. Notably, although the mean confidence between labeling tasks differed, the standard deviation of confidences range between 0.64 and 1.31, in other words, less than two Likert scale points. The difference in confidence between task and non-task was significant ($p < .001$) for the overall AP-type labels and both AP, and DA.[5]

If we again examine confidence scores with respect to each corpus, we also see a result similar to that for agreement values. That is, confidence is highest for the KVRET corpus and lowest for SCoSE, with the other task-oriented corpus being marginally higher than CABNC in most cases (Figure 8). For each label type, an ANOVA over confidence scores per-corpora concur with those of agreement. Overall results are significant ($p \leq .027$), and effect size is large for AP and AP-types ($\omega^2 > .14$), and medium for DA ($\omega^2 = .1$).[6] Post-hoc analysis shows the only place we see significant differences is

[5] Due to the small sample size of confidence scores (one score per-label) the resulting statistical power for AP and DA is .72, and .77, respectively.

[6] The resulting statistical power for DA is .72.

between KVRET and the other corpora, particularly with AP. Similarly, the difference between the two non-task-oriented corpora and bAbI is statistically non-significant in all cases. This indicates that, as with agreement, the division is not necessarily between task and non-task-oriented dialogs, but primarily between KVRET and the other three corpora.

These results show that there is a remarkable similarity between the annotators reported confidence scores and the resulting agreement values. When considered from the perspectives of task and non-task-oriented dialogs, individual corpora, and different label types, where higher confidence was reported, agreement was also higher. Annotators were therefore quite good at assessing how well their assigned labels fit the data, reporting higher confidence for dialogs where appropriate labels, or dialogue structure, was more intuitive, and lower confidence on the less structured dialogue types. This also suggests that incorporating confidence scores could be a valuable resource assessing labeling accuracy. Kazai (2011), showed that annotators who rated the task easier also had a higher accuracy. While Oyama et al. (2013), used self-reported confidence scores, along with their assigned labels, to estimate the "true" labels using the expectation-maximization (EM) algorithm.

Annotation time

The time participants took to completely annotate each utterance was also recorded. Because participants likely spent some time reading utterances and considering labels at the beginning of each dialogue, here all

Table 6. Mean and standard deviation of utterance annotation time (seconds) per corpus and dialogue type.

	KVRET	bAbI	Task	SCoSE	CABNC	Non-task
μ	24.62	33.57	29.09	25.56	36.69	31.13
σ	8.94	19.05	15.31	11.09	24.16	19.36

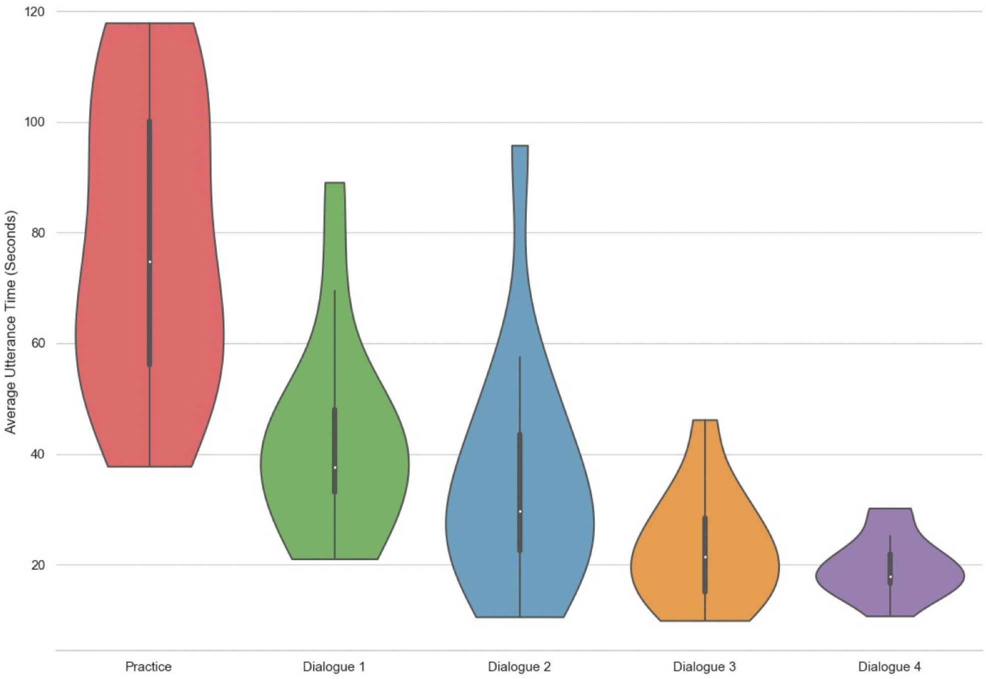


Figure 9. Distribution of annotators mean utterance annotation time (seconds) in the order dialogs were completed.

Table 7. Min, max, mean, and standard deviation of annotators mean utterance annotation time (seconds) in the order dialogs were completed.

Dialogue	Practice	1	2	3	4
Min	37.75	21.02	10.55	9.9	10.71
Max	117.87	89.06	95.76	46.17	30.20
μ	77.89	43.85	37.42	23.62	19.81
σ	27.52	20.48	24.85	11.08	6.03

reported times are the average time taken, in seconds, to annotate an utterance for that dialogue. Unlike agreement values and confidence scores, utterance times reveal that there is little difference between task and non-task-oriented dialogs, or the different corpora, as shown in Table 6. Therefore, despite reporting lower confidence for non-task-oriented dialogs, and the SCoSE corpus also containing around 4 times as many utterances, this did not seem to affect the average amount of time spent annotating those dialogs.

If we instead look at the average utterance time in the order dialogs were annotated, regardless of the specific dialogue, we see that annotation habits do indeed change over time. Figure 9 and Table 7 show that, for all participants, annotation time became faster as they progressed through the task, starting with an average of 77.89 seconds for the practice dialogue and ending with 19.81 seconds by dialogue 4. And further, that the variance between participants times also grew smaller over time, moving from a standard deviation of 27.52 on the practice dialogue, to just 6.03 on dialogue 4. These results seem to show a clear learning-effect, which echoes the results of Aulamo et al. (2019), where participants start with slow annotation speed, then, after a period of familiarization with the task, speed is increased and maintained for the remaining time. It may also be valuable to determine if there is a similar change in agreement over time, as annotators became more familiar with the schema and tool. Unfortunately, because all but the practice dialogue was shown in a random order for each participant, it is not possible to show that data and it will be left for future work. However, given that the practice dialogue also resulted in the highest agreement values, we suspect that this may not have a significant impact on agreement.

Conclusion

In this article, we have presented CAMS, which utilizes the CA concepts of AP, in conjunction with DA derived from the DiAML, to create a unified dialogue annotation scheme that captures the semantic and syntactic structure of dialogue for computational purposes. We assessed the schema by means of an exploratory annotation task, completed by novice annotators, and measured their inter-annotator agreement using dialogs from task-oriented and non-task-oriented settings. We also proposed distance functions, for each label type within the schema, that may be used when calculating inter-annotator agreement using weighted coefficients, such as Alpha and Beta.

Our findings indicate that inter-annotator agreement is significantly higher for the biased Beta coefficient, than that of unbiased Alpha, and this is principally caused by the differences in annotator label distributions increasing the Beta values. We therefore advise caution when comparing the two coefficients using the standard scales of interpretation (Geiß, 2021), particularly when biased measures diverge from unbiased ones. Nevertheless, if we assess agreement values of each dialogue set, using the somewhat arbitrary scale of Landis and Koch (1977), we find that for Beta DA and AP-type agreement can be considered "substantial," while AP fall into the "moderate" agreement category. However, agreement for the Alpha coefficient is less convincing. DA show a "moderate" level of agreement, while AP and AP-types only achieve "slight" and "fair" respectively. If we use the more stringent range [.67, .8], often used in Computational Linguistics to allow for "tentative conclusions to be drawn" (Carletta, 1996; Krippendorff, 2004), we find that only DA and AP-type labels are able to reach this threshold for the Beta coefficient. These results seem to concur with Poesio and Vieira (1998), and Hearst (1997), that reaching the .67 threshold is difficult for discourse annotation tasks. In this case, it may be due to our use of non-expert annotators, who have been shown to misunderstand the proper use of AP, and therefore more intense training should be provided, or expert annotators used. It may also be due to differences in individual annotator interpretations of the dialogs and appropriate AP labels. However, these agreement values can be considered an indication of moderate reliability.

Regarding task-oriented and non-task-oriented dialogs, both annotator agreement and self-reported annotator confidence scores are higher for task-oriented dialogs than non-task. However, when considered from the perspective of the individual corpora this distinction is not as clear. With the (task-oriented) KVRET corpus resulting in higher agreement and confidence scores than the other 3. We therefore conclude that, while CAMS is indeed applicable to both task and non-task-oriented dialogs, our results show that it is more intuitively applied to task-oriented dialogs. The determining factor, however, is not the division between task and non-task, but rather the content of the dialogue itself. Notably, we observed that utterances where the DA label is ambiguous, or multidimensional, can lead to different interpretations of the dialogue and result in a high number of disagreements for both DA *and* AP. Regarding the constituent label types within the schema, we found that DA labels consistently resulted in higher agreement and confidence scores than AP. This is perhaps not surprising, given that DA labels need only apply to one utterance at a time and generally use more intuitive names. AP on the other hand, require more specialized knowledge, and annotators must also consider relationships between utterances in order to apply them correctly. We found that many annotators misunderstood, and incorrectly applied the FPP and SPP labels. If labeling accuracy is required for the creation of an annotated corpus, this task may be better suited to experts, or novice annotators who have received more training than ours. Additionally, in order to produce accurate agreement scores the annotation tool intentionally placed no restrictions on label assignments; In future iterations this could be altered, to prevent, for example, the invalid creation of a new AP before a prior pair is completed. Unfortunately, given our procedural setup we were unable to measure if there is any improvement in agreement over time, once annotators had learned the annotation tool and schema. However, measuring the average time taken to annotate each utterance shows a clear pattern of learning, with annotation time decreasing for all annotators the longer they spent on the task. This indicates that the schema is inherently learnable and becomes more intuitive to apply with practice.

This article also explored some of the different assumptions around chance agreement for the unbiased (Alpha) and biased (Beta) agreement coefficients. We show, by means of JSD and Chi-squared analysis, that the annotators did indeed use similar distributions. Though the variance is larger for AP, which may require a greater degree of semantic interpretation, and where our annotators were often shown to misunderstand. However, these small differences in distributions resulted in dramatic differences between agreement scores for the Alpha and Beta coefficients, with consistently lower values for Alpha, and highlighting that the biased Beta coefficient is a less sensitive measure. Yet, if biased and unbiased measures diverge, the extent to which they do can provide useful information in its own right; by highlighting differences in annotator understanding of appropriate label categories, or between the annotation material itself. We therefore conclude that, if labeling accuracy is key, an unbiased measure such as Alpha should be used. However, for annotation tasks that require a high degree of semantic interpretation reporting both measures may be more beneficial.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

*Nathan Duran: Doctoral student in the Computer Science and Creative Technologies department at the University of the West of England. Research explores deep learning techniques for language representation and conversational interfaces inspired by human interaction.

*Steve Battle: Dr of Information Science in the Computer Science and Creative Technologies department at the University of the West of England. Research interests in applying Cybernetic principles to robotics, conversational interfaces, and the Internet of Things.

*Jim Smith: Professor in Interactive Artificial Intelligence (AI) and Deputy Director of the Computer Science Research Centre at the University of the West of England. Research interests in Machine Learning, Interactive AI, and Evolutionary Computation.

ORCID

Nathan Duran http://orcid.org/0000-0001-6084-4406
Steve Battle http://orcid.org/0000-0002-7154-7869
Jim Smith http://orcid.org/0000-0001-7908-1859

References

Albert, S., de Ruiter, L. E., & de Ruiter, J. (2015). *CABNC: The Jeffersonian transcription of the spoken British national corpus*. https://saulalbert.github.io/CABNC/

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog act markup in several layers* (Tech.Rep.).

Artstein, R., & Poesio, M. (2005a). Bias decreases in proportion to the number of annotators. In *Proceedings of the conference on formal grammar and mathematics of language (fg-mol)* (CSLI Publications) (pp. 141–150). http://web.stanford.edu/group/cslipublications/cslipublications/FG/2005/artstein.pdf

Artstein, R., & Poesio, M. (2005b, September). *Kappa 3 = Alpha (or Beta)* (Tech. Rep. No. (University of Essex)). Vol. 1. http://www.cs.pitt.edu/~wiebe/courses/CS3730/Fall08/poesioTechReportKappaCubed.pdf%5Cnpapers2://publication/uuid/F37A7D18-90E8-453B-9415-D0A821BF589D

Artstein, R., & Poesio, M. (2008). Inter-Coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Artstein, R. (2018). Inter-annotator agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). Springer.

Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., . . . Suleman, K. (2017). Frames: A corpus for adding memory to Goal-Oriented dialogue systems. In *Proceedings of the sigdial 2017 conference* (pp. 207–219). Saarbrucken, Germany: Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-5526

Aulamo, M., Creutz, M., & Sjoblom, E. (2019). Annotation of subtitle paraphrases using a new web tool. In *Proceedings of 4th conference of the association digital humanities in the nordic countries* CEUR-WS.org. CEUR-WS.org. http://urn.fi/urn:nbn:fi:

Austin, J. L. (1962). *How to do things with words*. Oxford University Press. http://pubman.mpdl.mpg.de/pubman/item/escidoc:2271128/component/escidoc:2271430/austin1962how-to-do-things-with-words.pdf

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725. https://doi.org/10.1162/COLI_a_00074

Bordes, A., Boureau, Y.-L., & Weston, J. (2017). *Learning End-to-EndGoal-Oriented Dialog* (ICLR 2017 (Association for Computational Linguistics)). https://arxiv.org/pdf/1605.07683.pdf.

Boxman-Shabtai, L. (2020). Meaning multiplicity across communication subfields: Bridging the gaps. *Journal of Communication*, *70* (3), 401–423. https://doi.org/10.1093/joc/jqaa008

Boyer, K. E., Ha, E. Y., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. (2009). Inferring tutorial dialogue structure with hidden Markov modeling. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications - edappsnlp '09* (Association for Computational Linguistics) (pp. 19–26). https://www.cs.rochester.edu/$\sim$tetreaul/bea4/Boyer-BEA4.pdfhttp://portal.acm.org/citation.cfm?doid=1609843.1609846

Boyer, K. E., Ha, E. Y., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. (2010). Dialogue act modeling in a complex task-oriented domain. In *Proceedings of sigdial 2010: the 11th annual meeting of the special interest group in discourse and dialogue* (Association for Computational Linguistics) (pp. 297–305).

British Standards Institution. (2012). *ISO 24617-2: Language resource management - Semantic annotation framework (SemAF) Part 2: Dialogue acts*. https://bsol-bsigroup-com

Bunt, H. (1978). *Conversational principles in question-answer dialogues*. Tubingen. pp. 119–142.

Bunt, H. (2006). Dimensions in dialogue act annotation. *Proceeding of LREC 2006* (: European Language Resources Association (ELRA)).

Bunt, H. (2011). The semantics of dialogue acts. In *International conference on computational semantics iwcs '11* (pp. 1–13). Oxford, England: Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0101http://aclweb.org/anthology/W/W11/W11-0101.pdf

Bunt, H. (2017). *Guidelines for using ISO standard 24617-2*. (Tech. rep). Tilburg Center for Cognition and Communication. https://dialogbank.uvt.nl/wpcontent/uploads/tdb/2015/12/ISO24617-2_Annotation_Guidelines2017.pdf.

Bunt, H. (2000, January). Dialogue pragmatics and context specification. In H. Bunt & W. Black (Eds.), *Abduction, belief and context in dialogue. Studies in computational pragmatics* (pp. 81–149). John Benjamins. https://doi.org/10.1075/nlp.1.03bun.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, *22*(2), 249–254 https://aclanthology.org/J96-2004/.

Chowdhury, S. A., Stepanov, E. A., & Riccardi, G. (2016). Transfer of corpus specific dialogue act annotation to ISO standard: Is it worth it? In *The international conference on language resources and evaluation* European Language Resources Association (ELRA) (Vol. 9, pp. 132–135). https://aclanthology.org/L16-1020/

Clift, R. (2016). *Conversation analysis*. Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Collins, H., Leonard-Clarke, W., & O'Mahoney, H. (2019). 'Um, er': How meaning varies between speech and its typed transcript. *Qualitative Research*, *19* (6), 653–668. https://doi.org/10.1177/1468794118816615

Craggs, R., & Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, *31*(3), 289–295. https://doi.org/10.1162/089120105774321109

Cuayahuitl, H., Yu, S., Williamson, A., & Carse, J. (2016). Deep reinforcement learning for multi-domain dialogue systems. In *Nips workshop on deep reinforcement learning* (pp. 1–9). Barcelona, Spain. https://arxiv.org/pdf/1611.08675.pdf

Di Eugenio, B., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. https://doi.org/10.1162/089120104773633402

Di Eugenio, B. (2000). On the usage of kappa to evaluate agreement on coding tasks. In *2nd international conference on language resources and evaluation, lrec 2000* (Barcelona, Spain: European Language Resources Association (ELRA)) (pp. 441–444).

Ekman, P., & Scherer, K. (1984). *Structures of social action - Studies in conversation analysis* (J. Atkinson & J. Heritage, Eds.). Cambridge University Press. http://ebooks.cambridge.org/ref/id/CBO9780511665868

Eric, M., & Manning, C. D. (2017). Key-Value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th annual sigdial meeting on discourse and dialogue* (Saarbrucken, Germany: Association for Computational Linguistics) (pp. 37–49). https://nlp.stanford.edu/blog/a-new-multi-turn-multi-

Firdaus, M., Golchha, H., Ekbal, A., & Bhattacharyya, P. (2020). A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cognitive Computation* (Springer Science,Business Media). https://doi.org/10.1007/s12559-020-09718-4

Ge, W., & Xu, B. (2015). Dialogue management based on multi-domain corpus. In *Annualmeeting of the special interest group on discourse and dialogue (sigdial)* (Prague, Czech: Republic Association for Computational Linguistics) (pp. 364–373). http://www.sigdial.org/workshops/conference16/proceedings/pdf/SIGDIAL48.pdf

Geertzen, J., & Bunt, H. (2010). Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th sigdial workshop on discourse and dialogue* (pp. 126–133). Sydney, Australia: Association for Computational Linguistics. http://ls0143.uvt.nl/dit/

Geertzen, J., Petukhova, V., & Bunt, H. (2008). Evaluating dialogue act tagging with naive and expert annotators. In *Proceedings of the 6th international conference on language resources and evaluation, lrec 2008* (pp. 1076–1082). Marrakech, Morocco: European Language Resources Association (ELRA).

Geiß, S. (2021). Statistical power in content analysis designs: How effect size, sample size and coding accuracy jointly affect hypothesis testing – A monte carlo simulation approach. *Computational Communication Research*, *3* (1), 61–89. https://doi.org/10.5117/ccr2021.1.003.geis

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press. http://www.getcited.org/pub/102129430

Green, J., Franquiz, M., & Dixon, C. (1997). The myth of the objective transcript: Transcribing as a situated act. *TESOL Quarterly*, *31* (1), 172. https://doi.org/10.2307/3587984

Griol, D., Hurtado, L., Segarra, E., & Sanchis, E. (2008). A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, *50*(8–9), 666–682. https://doi.org/10.1016/j.specom.2008.04.001

Grosz, B. J. (2018). Smart enough to talk with us? Foundations and challenges for dialogue capable ai systems. *Computational Linguistics*, *44*(1), 1–15. https://doi.org/10.1162/COLI_a_00313

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, *23* (1), 33–64. http://dl.acm.org/citation.cfm?id=972687%5Cnhttp://dl.acm.org/citation.cfm?id=972684.972687

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on Kappa n, Cohen's Kappa, Scott's π, and Aickin's α Understanding Statistics . *2*(3 p205–219 doi:10.1207/s15328031us0203_03).

Iseki, Y. (2019). Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *2019 22nd conference of the oriental cocosda international committee for the co-ordination and standardisation of speech databases and assessment techniques (o-cocosda)* (pp. 1–6). Cebu, Philippines: IEEE.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD-DAMSL ShallowDiscourse-function annotation coders manual* (Tech. Rep. (CU Boulder)). ftp://ftp.dcs.shef.ac.Uk/share/nlp/amities/bib/ics-tr-97-02.pdf

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. In *Proceed- ings ofthe 33rd european conference on information retrieval (ecir)* (Vol. 6611 Berlin, Heidelberg: LNCS, pp. 165–176). https://www.mturk.com/

Keizer, S., & Rieser, V. (2017). Towards learning transferable conversational skills using multi-dimensional dialogue modelling. In *Semdial 2017*. Saarbru¨cken, Germany (SEMDIAL).

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage Publications.

Kumar, V., Sridhar, R., Narayanan, S., & Bangalore, S. (2008). Enriching spoken language translation with dialog acts. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08(June)* (Columbus, Ohio: Association for Computational Linguistics), 225. http://www.aclweb.org/anthology/P08-2057http://portal.acm.org/citation.cfm?doid=1557690.1557755

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Li, X., Chen, Y.-N., Li, L., Gao, J., & Celikyilmaz, A. (2017). End-to-End Task-Completion neural dialogue systems. In *Proceedings of the the 8th international joint conference on natural language processing* (pp. 733–743). Taipei, Taiwan: AFNLP. http://aclweb.org/anthology/I17-1074http://arxiv.org/abs/1703.01008

Liddicoat, A. J. (2007). *An introduction to conversation analysis* (pp. 319). Continuum.

Macagno, F., & Bigi, S. (2018). Types of dialogue and pragmatic ambiguity Oswald, Steve and Herman, Thierry and Jacquin, Jerome. In *Argumentation and language-linguistic, cognitive and discursive explorations* (Vol. 32, pp. 191–218). Springer. isbn: 9783319739724. https://doi.org/10.1007/978-3-319-73972-4_9

Mezza, S., Cervone, A., Tortoreto, G., Stepanov, E. A., & Riccardi, G. (2018). ISO-Standard domain-independent dialogue act tagging for conversational agents. In *Coling 2018* (pp. 3539–3551). Santa Fe, New Mexico (Association for Computational Linguistics). http://arxiv.org/abs/1806.04327https://github.com/

Norrick, N. (2004). *Saarbrucken corpus of spoken English (SCoSE)*. https://ca.talkbank.org/access/SCoSE.html

Nowak, S., & Ru¨ger, S. (2010). How reliable are annotations via crowdsourcing? - A study about inter-annotator agreement for multi-label image annotation. In *Mir '10 proceedings of the international conference on multimedia information retrieval* (Philadelphia, Pennsylvania: Association for Computing Machinery) (p. 557). https://dl.acm.org/citation.cfm?id=1743478

Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013). Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Ijcai international joint conference on artificial intelligence* (Beijing, China: AAAI Press) (pp. 2554–2560).

Poesio, M., & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, *24*(2 183–216 doi: https://aclanthology.org/J98-2001/).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50* (1), 696–735. http://www.jstor.org/stable/412243http://about.jstor.org/terms

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*. Cambridge University Press.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19* (3), 321–325. https://www.jstor.org/stable/2746450

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H., & Hayward, C. S. U. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Sigdial 2004* (Berkeley CA: International Computer Science Inst) (pp. 97–100). https://aclanthology.info/pdf/W/W04/W04-2319.pdfhttp://www.aclweb.org/anthology/W04-2319

Sidnell, J. (2010). *Conversation analysis - An introduction*. Whiley-Blackwell. http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-40

Snow, R., Connor, B. O., Jurafsky, D., Ng, A. Y., Labs, D., & St, C. (2008). Cheap and fast - But is it good ? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 254–263). Honolulu: Association for Computational Linguistics. http://blog.doloreslabs.com/?p=109

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merrienboer, B., Joulin, A., & Mikolov, T. (2015). *Towards AI-Complete question answering: A set of prerequisite toy tasks. arXiv*. http://allenai.org/aristo.htmlhttp://arxiv.org/abs/1502.05698 ICLR

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). "Development and use of a gold standard data set for subjectivity classifications". In: *ACL '99: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* College Park, Maryland. ACM, pp. 246–253. https://doi.org/10.3115/1034678.1034721.

Williams, J. D., Raux, A., & Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue and Discourse*, *7* (3), 4–33. https://pdfs.semanticscholar.org/4ba3/39bd571585fadb1fb1d14ef902b6784f574f.pdf

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103* (3), 374–378. https://doi.org/10.1037/0033-2909.103.3.374