Dialogue Act Annotation with the ISO 24617-2 Standard

Harry Bunt, Volha Petukhova, David Traum and Jan Alexandersson

Abstract This chapter describes recent and ongoing annotation efforts using the ISO 24617-2 standard for dialogue act annotation. Experimental studies are reported on the annotation by human annotators and by annotation machines of some of the specific features of the ISO annotation scheme, such as its multidimensional annotation of communicative functions, the recognition of each of its nine dimensions, and the recognition of dialogue act qualifiers for certainty, conditionality, and sentiment. The construction of corpora of dialogues, annotated according to ISO 24617-2 is discussed, including the recent DBOX and DialogBank corpora.

1 Introduction

The ISO 24617-2 annotation standard has been designed for the annotation of spoken, written and multimodal dialogue with information about the dialogue acts that make up a dialogue, with the aim to create interoperable annotated resources. A dialogue act is a unit in the description of communicative behaviour that corresponds semantically to certain changes that the speaker wants to bring about in the information state of an addressee. ISO 24617-2 defines a dialogue act as:

(1) *Communicative activity of a dialogue participant, interpreted as having a certain communicative function and semantic content.*

The communicative function of a dialogue act, such as *Propositional Question, Inform, Confirmation, Request, Apology,* or *Answer*, specifies how the act's semantic content changes the information state of an addressee upon understanding the speaker's communicative behaviour.

According to the annotation schemes that existed prior to the establishment of ISO 24617-2 and its immediate predecessor DIT^{++}, such as DAMSL; MRDA; HCRC Map Task; Verbmobil; SWBD-DAMSL; and DIT[1], dialogue act annotation consisted of segmenting a dialogue into certain grammatical

[1] See Allen & Core (1997); Dhillon et al. (2004); Carletta et al. (1996); Jurafsky et al. (1997); Alexandersson et al. (1998); Bunt (1994; 2000)

units and marking up each unit with one or more communicative function labels. The ISO 24617-2 standard supports the annotation of dialogue acts in semantically more complete ways by additionally annotating the following aspects:

Dimensions The annotation scheme supports 'multidimensional' annotation, where multiple communicative functions may be assigned to dialogue segments; different from DAMSL and other multidimensional schemes, the ISO scheme uses an explicitly defined notion of 'dimension', which corresponds to a certain type of semantic content.

Qualifiers are defined for expressing that a dialogue act is performed conditionally, with uncertainty, or with a particular sentiment.

Functional and feedback dependence relations link a dialogue act to other units in a dialogue, e.g. for indicating which question is answered by a given answer, or which utterance a speaker is providing feedback about.

Rhetorical relations may optionally be annotated to indicate e.g. that one dialogue act motivates the performance of another dialogue act.

The following example illustrates the use of dimensions, communicative functions, qualifiers, dependence relations and rhetorical relations (where `"#fs1", "#fs2"`, and `"#fs3"` indicate the segments in P1's and P2's utterances that express a dialogue act - see Section 2.2 for more on segmentation).

(2) 1. P1: Is there an earlier connection?
2. P2: Ehm,.. no, unfortunately there isn't.

```
<diaml xmlns:"http://www.iso.org/diaml/">
<dialogueAct xml:id="da1" target="#fs1"
  sender="#p1" addressee="#p2"
  communicativeFunction="propositionalQuestion" dimension="task"/>
<dialogueAct xml:id="da2" target="#fs2"
  sender="#p2" addressee="#p1"
  communicativeFunction="stalling" dimension="timeManagement"/>
<dialogueAct xml:id="da3" target="#fs2"
  sender="#p2" addressee="#p1"
  communicativeFunction="turnTake" dimension="turnManagement"/>
<dialogueAct xml:id="da4" target="#fs3"
  sender="#p2" addressee="#p1"
  communicativeFunction="answer" dimension="task" sentiment="regret"
  functionalDependence="#da1"/>
</diaml>
```

The development of ISO 24617-2 was supported by annotation experiments in which preliminary versions of the scheme were tested for their usability by human annotators and by machine-learned annotation. After its establishment as an international standard in 2012, further annotation efforts have been undertaken in applying the standard in several corpus annotation, collection, and re-annotation projects. This chapter describes the most substantial of these experiments and annotation efforts.

This chapter is organized as follows. Section 2 outlines the use of the ISO 24617-2 annotation scheme. Section 3 describes the results of experiments concerned with some of the special features of the annotation scheme. Section 4 presents several new and emerging corpora of dialogues, annotated with the ISO 24617-2 annotation scheme. Section 5 closes this chapter with concluding remarks and perspectives for future studies and applications using the ISO 24617-2 standard.

2 Annotating with ISO 24617-2

2.1 Features of the ISO 25617-2 Annotation Standard

Dimensions. Utterances in dialogue often have more than one communicative function, as several authors have observed (Allwood, 1992; Bunt, 1994; 2011; Popescu-Belis, 2005; Traum, 2000). The following dialogue fragment illustrates this:

(3) 1. Anne: Henry, can you take us through these slides?
2. Henry: Ehm... sure, just ordering my notes.

In the first utterance, Anne makes a request and assigns the next speaking turn to Henry. In the second utterance, Henry accepts the turn and stalls for time; accepts the request, and explains why he does not fulfill the request right away. The multidimensional DIT^{++} annotation scheme was designed to optimally support the annotation of multifunctional utterances (Bunt, 2009). This scheme is based on a well-founded notion of dimension, inspired by the observation that participation in a dialogue involves a range of communicative activities beyond those strictly related to performing the task or activity that motivates the dialogue. Dialogue participants also perform communicative activities such as giving and eliciting feedback, taking turns, stalling for time, and showing attention; moreover, they often perform several of these activities at the same time. The term 'dimension' refers to these various types of communicative activity.

The ISO 24617-2 annotation scheme inherits the following nine dimensions from the DIT^{++} scheme: (1) *Task:* dialogue acts that move the task or activity forward which motivates the dialogue; (2-3) *Feedback*, divided into *Auto-* and *Allo-Feedback*: acts providing or eliciting information about the processing of previous utterances by the current speaker or by the current addressee, respectively; (4) *Turn Management:* activities for obtaining, keeping, releasing, or assigning the right to speak; (5) *Time Management:* acts for managing the use of time in the interaction; (6) *Discourse Structuring:* dialogue acts dealing with topic management, opening and closing (sub-)dialogues, or otherwise structuring the dialogue; (7-8) *Own-* and *Partner Communication Management:* actions by the speaker to edit his current contribution or a contribution of another current speaker, respectively; (9) *Social Obligations Management:*

dialogue acts for dealing with social conventions such as greeting, introducing oneself, apologizing, and thanking.

The ISO 224617-2 inventory of communicative functions consists of 56 of the 88 functions of the DIT^{++} taxonomy.[2] Some of these are specific for a particular dimension; for instance *Turn Take* is specific for Turn Management; Stalling is specific for Time Management, and *Self-Correction* is specific for Own Communication Management. Other functions can be applied in any dimension; for example, *You misunderstood me* is an *Inform* in the Allo-Feedback dimension. All types of question, statement, and answer can be used in any dimension, and the same is true for commissive and directive functions, such as *Offer, Suggest,* and *Request*. These functions are called *general-purpose* functions, as opposed to *dimension-specific* functions. Table 1 lists the communicative functions defined in ISO 24617-2.

Qualifiers. The different qualifiers defined in ISO 24617-2 are applicable to different classes of dialogue acts. Sentiment qualifiers are applicable to any dialogue act with a general-purpose function (GPF); conditionality qualifiers to dialogue acts with a commissive or directive function (*Promise, Offer, Suggestion, Request*, etc.); and certainty qualifiers are applicable to dialogue acts with an 'information-providing' function' (*Inform, Agreement, Disagreement, Correction, Answer, Confirm, Disconfirm).*

Functional dependence relations are indispensable for the interpretation of dialogue acts that are responsive in nature, such as *Answer, Confirmation, Disagreement, Accept Apology,* and *Decline Offer*. The semantic content of these acts depends crucially on the content of the dialogue act that they respond to. Functional dependence relations connect occurrences of such dialogue acts to their 'antecedent' and correspond to links for marking up a segment not only as having the function of an answer, for example, but also indicating which question is answered.

Feedback dependence relations play a similar role for determining the semantic content of feedback acts, which is co-determined by the utterance(s) that the feedback is about. Feedback acts often refer to the immediately preceding utterance, but can also refer further back and to more than one utterance (Petukhova, Prévot & Bunt, 2011). The ISO 24617-2 annotation scheme therefore includes links for marking up these dependences; an example occurs in (7).

Rhetorical relations, which have been studied extensively for written texts, also occur in spoken dialogue where they occur in two different ways, illustrated in the following examples (where the participants talk about remote TV controls):

[2] DIT^{++} has a fine-grained set of 29 feedback functions, whereas ISO 241617-2 has only 5, which are however more reliably annotated.

Table 1 ISO 24617-2 communicative functions

General-Purpose Communicative Functions	Dimension-Specific Communicative Functions	
	Function	Dimension
Inform	AutoPositive	Auto-Feedback
Agreement	AutoNegative	
Disagreement	AlloPositive	Allo-Feedback
Correction	AlloNegative	
Answer	FeedbackElicitation	
Confirm	Staling	Time Management
Disconfirm	Pausing	
Question	Turn Take	Turn Management
Set-Question	Turn Grab	
Propositional Question	Turn Accept	
Choice-Question	Turn Keep	
Check-Question	Turn Give	
Offer	Turn Release	
Address Offer	Self-Correction	Own Communication Man.
Accept Offer	Self-Error	
Decline Offer	Retraction	
Promise	Completion	Partner Communication Man.
Request	Correct Misspeaking	
Address Request	Interaction Structuring	Discourse Structuring
Accept Request	Opening	
Decline Request	Init-Greeting	Social Obligations Man.
Suggest	Return Greeting	
Address Suggest	Init-Self-Introduction	
Accept Suggest	Return Self-Introduction	
Decline Suggest	Apology	
Instruct	Accept Apology	
	Thanking	
	Accept Thanking	
	Init-Goodbye	
	Return Goodbye	

(4) 1. A: I can never find them.
2. B That's because they don't have a fixed location.

(5) 1. A: Where would you position the buttons?
2. A: I think that has some impact on many things

In (4) the dialogue acts expressed by A's and B's utterances are related by a *Cause* relation between their respective semantic contents: the content of the second causes the content of the first; in (5), by contrast, the second dialogue act forms a reason for performing the first, so the causal relation is between the two dialogue acts as a whole, rather than between their semantic contents. The annotation of a rhetorical relation is illustrated in example (8).

Different from functional and feedback dependences, which are an integral part of dialogue acts with a responsive function and of feedback acts,

respectively, rhetorical relations give additional information about the ways in which dialogue acts are semantically or pragmatically related.

2.2 Multidimensional Segmentation

Dialogues are often segmented into *turns*, defined as stretches of communicative behaviour produced by one speaker, bounded by periods of inactivity of that speaker. Such a segmentation is too coarse for accurate dialogue act annotation, as example (3) above illustrates. More accurate annotation is possible by using *'functional segments'* as the units to which annotations are attached. Functional segments are defined as the *minimal stretches of communicative behaviour that have a communicative function* - 'minimal' in the sense of not containing material that does not contribute to its communicative function(s). Functional segments are mostly shorter than turns, may be discontinuous, may overlap, and may have parts contributed by different speakers. Functional segments by definition have *at least one* communicative function, and possibly several. An example of the use of functional segments is shown in (6), where we see the utterance *The first train to the airport on Sunday is at...let me see... 6.16* in response to the the question *What time is the first train to the airport on Sunday?* The response has parts which have a communicative function in three different dimensions: Task, Auto-Feedback (expressed by the repetition in the second utterance), and Time Management; in each of these dimensions the relevant functional segment is shown; the DiAML annotation is represented in (7).

(6) C: What time is the first train to the airport on Sunday?
I: The first train to the airport on Sunday is at...let me see... 6.16

Auto-Feedback	fs2	*The first train to the airport on Sunday*
Task:	fs3	*The first train to the airport on Sunday is at 6.16*
Time Man.	fs4	*...let me see...*

(7)
```
<diaml xmlns:"http://www.iso.org/diaml/">
<dialogueAct xml:id="da1" target="#fs1"
  sender="#p1" addressee="#p2"
  communicativeFunction="setQuestion" dimension="task"/>
<dialogueAct xml:id="da2" target="#fs2"
  sender="#p2" addressee="#p1" communicativeFunction="autoPositive"
  dimension="autoFeedback" feedbackDependence="#fs1"/>
<dialogueAct xml:id="da3" target="#fs3"
  sender="#p2" addressee="#p1" communicativeFunction="answer"
  dimension="task" functionalDependence="#da1"/>
<dialogueAct xml:id="da4" target="#fs4"
  sender="#p2" addressee="#p1"
  communicativeFunction="stalling" dimension="timeManagement"/>
</diaml>
```

2.3 The Dialogue Act Markup Language (DiAML)

The ISO 24617-2 standard includes the specification of the Dialogue Act Markup Language (DiAML), designed in accordance with the ISO Linguistic Annotation Framework (ISO 24612, see ISO (2011), which draws a distinction between the concepts of *annotation* and *representation*. The term 'annotation' refers to the linguistic information that is added to segments of language data, independent of the format in which the information is represented; 'representation' refers to the format in which an annotation is rendered, independent of its content (Ide & Romary, 2004).

This distinction is implemented in the DiAML definition following the ISO Principles for Semantic Annotation (ISO 24617-6 (ISO, 2016); see also Bunt, 2015). The definition specifies, besides a class of XML-based *representation structures*, also a class of more abstract *annotation structures* with a formal semantics. These components are called the *concrete* and *abstract syntax*, respectively. Annotation structures are set-theoretical structures, like pairs and triples, for which the concrete syntax defines an XML-based rendering. An annotation structure is a set of *entity structures*, which contain semantic information about a functional segment, and *link structures*, which describe semantic relations between functional segments. An entity structure contains the conceptual information of a single dialogue act, and specifies: (1) a sender; (2) one or more addressees; (3) possible other participants, like an audience or side-participants; (4) a communicative function; (5) a dimension; (6) possible qualifiers for sentiment[3], conditionality or certainty; and (7) zero, one or more functional dependence relations or feedback dependence relations.

The concrete syntax, defined following the CASCADES method (see ISO 24617-6 and Bunt, 2015), has a unit that corresponds to entity structures in the form of the XML element `dialogueAct`, as illustrated in (2). The question asked by participant P1 is represented by the `dialogueAct` element with identifier `da1`, which refers to the functional segment `fs1` formed by P1's utterance. Participant P2's response consists of two functional segments. First, a turn-initial *Ehm,...* which forms a multifunctional segment signalling that P2 is taking the turn and also stalls for time. The second functional segment contains the actual answer, which includes an expression of regret that is annotated by means of a qualifier, represented as the value of the `sentiment` attribute.

Functional dependence relations are components of a `dialogueAct` element since they form part of a dialogue act viewed as a semantic unit. The same is true for feedback dependence relations as a component of a feedback act, as illustrated in example (7). Rhetorical relations, by contrast, do not play a role in determining the meaning of a dialogue act, but provide additional information about the semantic/pragmatic relations between dialogue

[3] ISO 24617-2 does not prescribe the use of any particular set of sentiment labels. See e.g. the EmotionML language (`www.w3.org/TR/emotionml`) for possible choices in this respect.

acts. They are represented by means of `rhetoricalLink` elements as shown in (8).

(8) 1. P4: Where would you position the buttons?
2. P4: I think that has some impact on many things

```
<diaml xmlns:"http://www.iso.org/diaml/">
<dialogueAct xml:id="da1" target="#fs1"
  sender="#p4" addressee="#p3"
  communicativeFunction="setQuestion" dimension="task"/>
<dialogueAct xml:id="da2" target="#fs2"
  sender="#p4" addressee="#p3"
  communicativeFunction="inform" dimension="task"/>
<rhetoricalLink dact="#da2"
  rhetoRelatum="#da1" rhetoRel="cause"/>
</diaml>
```

3 Experiences in the Use of ISO 24617-2

3.1 Communicative Function Recognition

Multidimensional annotation using a rich inventory of dialogue act tags is often thought to be too difficult for human annotators as well as for automatic annotation to give reliable results. In order to investigate this, Geertzen & Bunt (2006) determined the inter-annotator agreement for assigning communicative functions in the ten dimensions of DIT^{++}, nine of which are inherited by ISO 24617-2.

They observed that, when a hierarchically structured tag set is used, the popular standard kappa coefficient (Cohen, 1960) is not an appropriate measure of agreement, since the assignment to a functional segment of two different but hierarchically related tags, like *Answer* and *Confirm*, or *Inform* and *Agreement*, does not reflect total disagreement, as the standard kappa would assume, but *partial* (dis-)agreement, since a *Confirm* act is a particular kind of *Answer*, and an *Agreement* is a particular kind of *Inform*. Instead, they defined a weighted kappa coefficient, using Cohen's weighted kappa coefficient (Cohen, 1968) with a distance metric that takes the hierarchical structure of the tag set into account (see also Lesch et al., 2005). The *taxonomically weighted kappa* is defined as follows:

(9) $\kappa_{tw} = 1 - \frac{\Sigma(1-\delta(i,j)).P_{oi}j}{\Sigma(1-\delta(i,j)).P_{ei}j}$

where the distance metric δ_{ij} measures disagreement and is a real number normalized in the range between 0 and 1 (P_{oi} and P_{ei} are observed and expected probabilities, respectively). Table 2 shows standard and taxonomically weighted kappa scores per ISO 246170-2 dimension, averaged over all annotation pairs, for the DIAMOND corpus[4].

[4] See Geertzen et al. (2004).

Table 2 Standard and weighted kappa-scores for annotator agreement in the annotation of communicative functions, per ISO 24617-2 dimension (adapted from Geertzen & Bunt, 2006).

Dimension	standard kappa			weighted kappa		
	P_o	P_e	κ	P_o	P_e	κ_{tw}
Task	0.52	0.09	0.47	0.76	0.17	0.71
Auto-Feedback	0.32	0.14	0.21	0.87	0.69	0.57
Allo-Feedback	0.53	0.19	0.42	0.79	0.50	0.58
Turn Management	0.90	0.42	0.82	0.90	0.42	0.82
Time Management	0.91	0.79	0.58	0.91	0.79	0.58
Own Communication Management	1.00	0.50	1.00	1.0 0	0.95	1.00
Partner Communication Management	1.00	1.00	–	1.00	1.00	–
Dialogue structuring	0.87	.0.48 0.	74	0.87	0.48	0.74
Social Obligation Management	1.00	0.19	1.00	1.00	0.19	1.00

The agreement scores indicate that human annotators can reliably use a rich, multidimensional annotation scheme like ISO 24617-2 or DIT^{++}. The usability and reliability of an annotation scheme is not just a matter of the size or simplicity of the tag set, but rather of the conceptual clarity of the tags, their definitions and accompanying annotation guidelines.

3.2 Dimension Recognition

The notion of a dimension, as used in ISO 24617-2 and DIT^{++}, is defined as follows (Bunt, 2004).:

(10) *A dimension is a class of dialogue acts concerned with one particular aspect of communication that a dialogue act can address independently from other aspects*

Geertzen et al. (2008) assessed the recognizability of dimensions by human annotators and by automatic means. Three annotators independently annotated dialogues from the DIAMOND and OVIS[5] corpora with dimension tags. Table 3 presents agreement scores expressed in terms of Cohen's kappa and tagging accuracy (comparing with a gold standard, see Geertzen et al., 2008). The table shows near perfect agreement between annotators, and moreover that accuracy is very high. Human annotators can apparently recognize the dimensions of the ISO 24617-2 standard almost perfectly.

To assess the machine learnability of dimension recognition, the rule induction algorithm Ripper was applied to data from the AMI, OVIS, and DIAMOND corpora. The features included in the data sets relate to prosody (minimum, maximum, mean, and standard deviation of pitch); energy; voic-

[5] See `http://www.let.rug.nl/vannoord/Ovis/`.

Table 3 Inter-annotator agreement and tagging accuracy per dimension for the OVIS and DIAMOND corpora.

Dimension	Annotator agreement			Accuracy		
	P_o	P_e	κ	P_o	P_e	κ
Task	0.85	0.1	0.83	0.91	0.47	0.81
Auto-Feedback	0.91	0.1	0.90	0.94	0.24	0.92
Allo-Feedback	0.93	0.1	0.92	0.95	0.43	0.91
Turn Management	0.93	0.1	0.92	0.92	0.08	0.92
Time Management	0.99	0.1	0.99	0.99	0.11	0.90
Discourse Structuring	0.99	0.1	0.99	0.87	0.05	0.87
Contact Management	0.99	0.1	0.99	0.91	0.14	0.89
Own Communication Man.	0.99	0.1	0.99	1.00	0.02	1.00
Partner Communication Man.	0.99	0.1	0.99	1.00	0.02	1.00
Social Obligation Man.	0.99	0.1	0.99	0.95	0.09	0.95

ing; duration; occurrence of words (a bag-of-words vector); and dialogue history: tags of 10 previous turns. Table 4 presents the scores obtained in 10-fold cross-validation experiments. The results indicate that the dimensions of DIT^{++} and ISO 24617-2 are automatically recognizable with fairly high accuracy.

Table 4 Automatic dimension recognition scores in terms of accuracy (in %), with baseline scores (BL, classifier based on the dimension tag of the previous utterance), for AMI, DIAMOND, and OVIS data sets.

Dimension	DIAMOND		AMI		OVIS	
	BL	Accuracy	BL	Accuracy	BL	Accuracy
Task	64.9	70.5	66.8	72.3	60.8	73.5
Auto-Feedback	71.1	85.1	77.9	89.7	66.1	75.9
Allo-Feedback	86.9	96.6	96.7	99.3	52.5	80.1
Turn Management	69.5	90.0	59.0	93.0	89.8	99.2
Time Management	65.6	82.2	69.7	99.4	95.5	99.4
Discourse Structuring	59.0	67.9	98.0	92.5	76.3	89.4
Contact Management	88.0	95.2	99.8	99.8	87.7	98.5
Own Communication Man.	77.4	83.1	89.6	94.1	99.7	99.7
Partner Communication Man.	45.4	62.6	99.7	99.7	99.8	99.8
Social Obligation Management	80.3	92.2	99.6	99.6	96.2	98.4

3.3 Machine-learned Dialogue Act Recognition

Petukhova and Bunt (2011) investigated the automatic classification of dialogue acts for unsegmented spoken dialogue. Table 5 shows the results of the combined classification of dimension and communicative function, using three different 'local' classifiers that apply to local utterance features. The DER_{sc}

error-rate metric is based on the Dialog Act Error Rate (DER) defined by Zimmermann et al. (2005), which considers a word to be correctly classified if it has been assigned the correct dialogue act type, and it lies in the correct segment. Table 6 shows the results for two-step classification (manual segmentation followed by communicative function classification), which can be seen to work better for all dimensions except the Task dimension (the most important one).

Table 5 Overview of F- and DER_{sc}-scores for joint segmentation and classification in each ISO 24617-2 dimension for Map Task data. (Best scores in bold face.)

Classification task	BL		BayesNet		Ripper	
Dimension	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}
Task	43.8	70.2	**79.7**	41.9	77.7	58.5
Auto-Feedback	64.6	60.6	65.4	55.2	**80.1**	43.9
Allo-Feedback	30.7	91.2	59.3	54.0	**72.7**	51.8
Turn Management	50.3	47.5	70.8	40.9	**81.4**	36.2
Time management	54.2	28.4	72.1	20.3	**83.6**	10.4
Discourse Structuring	33.2	95.1	62.5	44.3	**66.7**	43.5
Contact Management	24.7	93.2	**57.0**	79.5	11.0	93.5
Own Communication Man.	11.2	97.4	**42.9**	64.7	28.6	92.1
Partner Communication Man.	14.3	95.2	61.5	55.2	**66.7**	50.1
Social Obligations Management	08.8	96.2	40.0	71.8	**85.7**	21.4

Table 6 Overview of F-scores on baseline (BL) and classifiers for two-step segmentation and classification tasks. (Best scores in bold face.)

Classification	BL	NBayes	Ripper	IB1
Task	66.8	71.2	**72.3**	53.6
Auto-Feedback	77.9	86.0	**89.7**	85.9
Allo-Feedback	79.7	**99.3**	99.2	98.8
Turn M.: initial	93.2	92.9	93.2	88.0
Turn M.: final	58.9	85.1	**91.1**	69.6
Time management	69.7	99.2	**99.4**	99.5
Discourse Structuring	69.3	**99.3**	**99.3**	99.1
Contact Management	89.8	99.8	99.8	99.8
Own Communication Management	89.6	90.0	**94.1**	85.6
Partner Communication Management	99.7	99.7	99.7	99.7
Social Obligations Management	99.6	99.6	99.6	99.6

The fact that dialogue utterances are often multifunctional, having a communicative function in more than one dimension, makes dialogue act recognition a complex task. Splitting up the task may make it more manageable. A widely used strategy is to split a multi-class learning task into several binary learning tasks. Learning multiple classes, however, allows a learning algorithm to exploit interactions among classes. Petukhova and Bunt (2011)

split the task in such a way that a classifier needs to learn (1) communicative functions in isolation; (2) semantically related functions together, e.g. all information-seeking functions (all types of questions) or all information-providing functions (all types of answers and informs). In total 64 classifiers were built for dialogue act recognition in AMI data and 43 for Map Task data.

Using local classifiers that produce all possible output predictions ('hypotheses') given a certain input leads to some predictions being false, since a local classifier never revisits a decision that it has made, in contrast with a human interpreter. Decisions should preferably be based not only on local features of the input, but also on broader contextual information. Therefore, Petukhova and Bunt (2011) trained higher-level 'global' classifiers that have, along with features extracted locally from the input data, the partial output predicted so far from all local classifiers. (This technique is also called 'meta-classification' or 'late fusion'). Five previously predicted class labels were used, taking into account that the average length of a functional segment in the data is 4.4 tokens. This was found to result in a 10-15% improvement. Some incorrect predictions are still made, since the decision is sometimes based on incorrect previous predictions.

A strategy to optimize the use of output hypotheses is to perform a global search in the output space looking for best predictions. This is not always the best strategy, however, since the highest-ranking predictions are not always correct in a given context. A possible solution is to postpone decision until some (or all) future predictions have been made for the rest of the current segment. For training, the classifier then uses not only previous predictions as additional features, but also future predictions of local classifiers. This forces the classifier to not immediately select the highest-ranking predictions, but to also consider lower-ranking predictions that could be better in the context.

Table 7 gives an overview of the global classification results based on added previous and next predictions of local classifiers. Both classifiers performed very well, outperforming the use of only local classifiers by a broad margin (cf. Table 5). It may be noted that the overall performance reported here is substantially better than the results of other approaches that have been reported in the literature. For instance, Reithinger and Klesen (1997) report a average tagging accuracy of 74.7% of applying techniques based on n-gram modelling to Verbmobil data; transformation-based learning applied to the same data achieved an accuracy of 75.1% (Samuel et al., 1998). Hidden Markov Models used for dialogue act classification in the Switchboard corpus gave a tagging accuracy of 71% (Stolcke et al., 2000); and Lendvai et al. (2004) report an accuracy of 73.8% for the application to data from the OVIS corpus of a memory-based approach based on the k-nearest neighbour algorithm.

Altogether, an incremental, token-based approach with global classifiers that exploit the outputs of local classifiers, applied to previous and subsequent tokens, results in excellent dialogue act recognition scores for unseg-

Table 7 Overview of F-scores and DER_{sc} when global classifiers are used for AMI and Map Task data, based on added predictions of local classifiers for five previous and five next tokens. (Best scores in bold face.)

Classification	AMI data				Map Task data			
	BayesNet		Ripper		BayesNet		Ripper	
Dimension	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}
Task	82.6	9.5	**86.1**	8.3	**85.8**	12.2	80.8	9.1
Auto-Feedback	81.9	1.9	**95.1**	0.6	84.4	15.0	**93.0**	7.6
Allo-Feedback	**96.3**	0.6	95.7	0.5	**95.3**	4.6	94.6	6.9
Turn Management:initial	**85.7**	1.5	81.5	1.6	89.5	8.2	**91.0**	8.0
Turn Management:close	90.9	3.8	**91.2**	3.6	**82.9**	17.1	77.2	18.9
Time management	90.4	2.4	**93.4**	1.7	**94.9**	5.5	92.8	6.1
Discourse Structuring	**82.1**	1.7	78.3	1.8	85.7	12.4	**87.4**	8.2
Contact Management	87.9	1.2	**94.3**	0.6	87.4	9.9	**88.3**	7.4
Own Communication Man.	78.4	2.2	**81.6**	2.0	87.2	9.8	**87.4**	7.6
Partner Communivation Man.	**71.8**	2.4	70.0	4.6	86.7	11.1	**86.8**	9.8
Social Obligations Man.	98.6	0.4	98.6	0.5	**97.9**	1.1	**97.9**	1.2

mented spoken dialogue. This can be seen as strong evidence for the machine learnability of the ISO 24717-2 annotation scheme.

3.4 Qualifier Recognition

The recognition of dialogue act qualifiers by human annotators was investigated by Petukhova (2011). The task in these experiments, involving four untrained annotators (undergraduate students), was to assign qualifier values to functional segments in pre-annotated dialogue fragments from the AMI corpus and the TRAINS corpus.[6]

Table 8 shows that there are no systematic differences between annotators in assigning values for qualifier tags. They achieved moderate agreement ($0.4 < \kappa < 0.6$) on labelling certainty for the AMI data; the agreement for this category when labelling TRAINS dialogues is substantial ($0.6 < \kappa < 0.8$). The difference can be explained by the fact that AMI dialogues are more difficult to annotate for untrained annotators: AMI meetings are considerably more complex, as they are both multi-party and multi-modal. The best recognized category is conditionality, for which annotators achieved substantial to near perfect agreement ($\kappa > 0.8$).

Inter-annotator agreement scores for certainty and sentiment were influenced negatively by the fact that one of the values that annotators could choose for these qualifiers was 'neutral'; some annotators assigned this qualifier to every segment that did not clearly express a certainty or a sentiment, while others assigned a certainty or a sentiment qualifier only to

[6] See `https://www.cs.rochester.edu/research/speech`.

Table 8 Cohen's kappa scores for inter-annotator agreement on the assignment of qualifiers per annotator pair for AMI and TRAINS data.

Annotator	AMI dialogues			TRAINS dialogues	
pair	Certainty	Conditionality	Sentiment	Certainty	Conditionality
1, 2	0.49	0.79	0.70	0.64	0.88
1, 3	0.48	0.64	0.66	0.70	0.73
1, 4	0.42	0.65	0.25	0.64	0.93
2, 3	0.47	0.85	0.60	0.68	0.64
2, 4	0.35	0.79	0.36	0.71	0.88
3, 4	0.38	0.65	0.30	0.75	0.73

those segments which they judged as expressing a particular sentiment or (un)certainty.

4 Annotated Corpora

4.1 The DBOX Corpus

In the European project DBOX[7], which aims to develop interactive games based on spoken natural language human-computer dialogues, a corpus has been collected in a Wizard-of-Oz setting. A set of quiz games was designed where the Wizard holds the facts about a famous person's life and the player's task is to guess this person's name by asking questions.

In total 338 dialogues were collected with a total duration of 16 hours, comprising about 6,000 speaking turns. The collected data has been transcribed and annotated using the ISO 24617-2 annotation scheme. Table 9 shows that inter-annotator agreement between two trained annotators ranged between 0.55 and 0.94 in terms of Cohen's kappa for segmentation and between 0.55 and 1.00 for the annotation of dialogue acts in the various dimensions (see Petukhova et al., (2014) for details). For relations between dialogue acts the agreements ranged from 0.66 to 0.88.

[7] Eureka project E! 7152, see `https://www.lsv.uni-saarland.de/index.php?id=71`.

Table 9 Inter-annotator agreement on segmentation and annotation of communicative functions per ISO dimension and on annotation of relations of the ISO relation types.

ISO 24617-2 dimension	segmentation (κ)	function (κ)
Task	0.88	0.81
Auto-feedback	0.78	0.79
Allo-Feedback	0.94	0.95
Turn Management	0.71	0.64
Time Management	0.86	0.86
Discourse Structuring	0.88	0.55
Own Communication Management	0.55	0.98
Partner Communication Management	n.a.	n.a.
Social Obligations Management	0.77	1.00
ISO 24617-2 relation type		relations
Functional dependence	0.88	0.68
Feedback dependence	0.88	0.88
Rhetorical relations	0.88	0.68

4.2 Youth Parliament Debate Data

As part of the FP 7 European project Metalogue[8], data have been analysed from three sessions of the UK Youth Parliament (YP). The sessions are video recorded and available on YouTube[9]

The annotated corpus consists of 1388 functional segments from 35 speakers. Table 10 provides an overview of the relative frequencies of functional tags per ISO-dimension.

Table 10 Distribution of functional tags across ISO-dimensions in the UK YP corpus.

ISO 24617-2 Dimension	Frequency
Task	54.9 %
Auto Feedback	2.9 %
Allo Feedback	1.0 %
Turn Management	22.7 %
Time Management	21.1 %
Discourse Structuring	10.0 %
Own Communication Management	7.3 %
Partner Communication Management	0.0 %
Social Obligations Management	1.2 %

Of the dialogue acts in the Task dimension, 41.4% are *Inform* acts, which are often connected by rhetorical relations. For example:

[8] See `http://www.metalogue.eu`.

[9] See for example `http://www.youtube.com/watch?v=g2Fg-LJHPA4`. For information about the UK Youth Parliament see `http://www.ukyouthparliament.org.uk/`.

(11)
$D1_{21}$: Let us be clear, sex education covers a wide range of issues affecting young people [*Inform*]
$D1_{22}$: These include safe sex practices, STIs and legal issues surrounding consent and abuse [*Inform Elaboration* $D1_{21}$]

The ISO 24617-2 standard does not prescribe the use of any particular set of rhetorical relations; for the annotation of the DBOX corpus a combination was used of the hierarchy of relations used in the Penn Discourse Treebank (PDTB, Prasad et al., 2008) and the taxonomy defined in Hovy and Maier (1995). Table 11 shows the distribution in the corpus of the rhetorical relations associated with *Inform* acts. The corpus is used for designing the Dialogue Manager module of the dialogue system that is built in the Metalogue project.

Table 11 Distribution of rhetorical relations associated with Inform acts in the corpus (*= as defined in the PDTB; **= as defined by Hovy and Maier, 1995; ***= in both taxonomies); inter-annotator agreement in terms of Cohen's kappa.

Rhetorical relation	relative frequency	annotator agreement
Elaboration**	28.1	0.67
Evidence**	21.4	0.72
Justify***	16.1	0.76
Condition***	0.7	0.34
Motivation**	1.4	0.48
Background**	0.3	0.18
Cause***	3.4	0.37
Result***	2.2	0.26
Reason*	10.6	0.33
Conclude**	5.7	0.71
Restatement***	10.1	0.76

4.3 The SWBD-ISO Corpus

Fang and collaborators made an effort to assign ISO 24617-2 annotations to the dialogues in the Switchboard Dialog Act (SWBD-DA) corpus (see Fang et al., 2011; 2012a; 2012b; Bunt et al. 2013)).[10] This resource contains 1,155 5-minute conversations, orthographically transcribed in about 1.5 million word tokens. Each utterance in the corpus is segmented in 'slash units', defined as *"maximally a sentence; slash units below the sentence level correspond to parts of the narrative which are not sentential but which the annotator interprets as complete"* (Meteer and Taylor 1995). The corpus comprises 223,606

[10] The Switchboard corpus is distributed by the Linguistic Data Consortium: `https://www.ldc.upenn.edu`.

slash units, which are annotated with a communicative function tag from the SWBD-DAMSL annotation scheme, a variation of the DAMSL scheme defined specifically for this purpose (Jurafsky et al. 1997). See example (12), where 'qy' is the SWBD-DAMSL tag for yes/no questions and 'utt1' indicates the first slash unit within a turn.

(12) qy A.1 utt1: { D Well, } { F uh, } does the company you work for test for drugs? /

In addition to this marking up of communicative functions, in-line markups are also used to mark 'discourse markers' such as { D Well, }, which often signal a rhetorical relation; filled pauses, like { F uh, }, restarts and repetitions, such as [I think, I think] and some other types of 'disfluencies'.

To assess the possibility of converting SWBD-DA annotations to ISO 24617-2 annotations, first a detailed comparison was made of the two sets of communicative functions, revealing 14 one-to-one correspondences and 26 many-to-one equivalences. These tags can thus be converted automatically to ISO tags, which accounts for 83,97% of the SWBD-DAMSL tags in the corpus. Six SWBD-DAMSL function tags have a one-to-many correspondence with 26 ISO tags, corresponding to 5.74% of the Switchboard corpus; about 30% of these cases can be converted automatically to an ISO tag by taking the tagging of the preceding slash unit into account; for example, an utterance tagged 'aa' (i.e. Accept) following an offer should be assigned the ISO tag *Accept Offer*, while it should be assigned the ISO tag *Accept Request* when following a request. For those cases where such a contextual disambiguation does not help, manual annotation was performed (see Fang et al., 2012b).[11]

Altogether, through combined automatic conversion and manual annotation 200.605 utterances (89.71% of the Switchboard corpus) were assigned ISO 24617-2 communicative function tags. Table 12 shows the distribution of function tags in the resulting 'SWBD-ISO' corpus.

4.4 The DialogBank

In a recent initiative at Tilburg University a publicly available corpus has been created called the **DialogBank**, which consists of dialogues with gold standard annotations in DiAML according to the ISO 24617-2 standard. While recommending the use of XML for representing annotation structures as defined by the DiAML abstract syntax, the standard allows representations in other formats as long as these have the properties of being (1) *complete*, i.e. defining a rendering of any annotation structure defined by the abstract syntax, and (2) *unambiguous*, i.e. every representation encodes only one annotation structure. Representation formats that have these properties can

[11] The remaining 10.29% of SWBD-DAMSL tags cannot be converted into ISO tags since they are not really concerned with communicative functions, such as the SWBD-DAMSL tags 'non-verbal', 'uninterpretable', 'quoted material', 'transcription error'.

Table 12 Distribution of ISO 24617-2 communicative function tags the SWBD-ISO corpus

ISO 24617-2	Utterances			ISO 24617-2	Utterances		
Comm. Functions	#	%	Cum %	Comm. Functions	#	% Cum %	
inform	120227	53.767	53.77	instruct	106	0.047	89.44
autoPositive	46382	20.743	74.51	acceptSuggest	99	0.044	89.48
agreement	10934	4.890	79.40	acceptApology	79	0.035	89.52
propositionalQuestion	5896	2.637	82.04	thanking	79	0.035	89.55
confirm	3115	1.393	83.43	offer	71	0.032	89.58
initialGoodbye	2661	1.190	84.62	acceptRequest	65	0.029	89.61
setQuestion	2174	0.972	85.59	signalSpeakingError	56	0.025	89.64
disconfirm	1597	0.714	86.31	promise	41	0.018	89.66
answer	1522	0.681	86.99	correction	29	0.013	89.67
checkQuestion	1471	0.658	87.64	acceptOffer	26	0.012	89.68
completion	813	0.364	88.01	turnTake	18	0.008	89.69
question	680	0.304	88.31	alloPositive	17	0.008	8970
stalling	580	0.259	88.57	correctMisspeaking	14	0.006	89.70
choiceQuestion	506	0.226	88.80	selfCorrection	8	0.004	89.71
suggest	369	0.165	88.96	acceptThanking	6	0.003	89.71
autoNegative	307	0.137	89.10	declineOffer	3	0.001	89.71
request	278	0.124	89.22	declineRequest	3	0.001	89.71
disagreement	258	0.115	89.34	turnRelease	2	0.001	89.71
apology	112	0.050	89.39	declineSuggest	1	0.000	89.71
non-functional tags					23001	10.29	100.00
Total					223606	100.00	

be converted to and from the DiAML-XML format without loss of information. For some of the dialogues in the DialogBank, an alternative tabular representation format was defined that has these properties and that is more convenient for human readers (see Bunt et al., 2016).

The annotations include not only the multidimensional marking up of communicative functions and dimensions, but also of functional dependence relations; feedback dependence relations; rhetorical relations; and qualifiers for certainty, conditionality and sentiment.

The DialogBank currently contains dialogues taken from four English-language corpora: the HCRC Map Task, Switchboard, TRAINS, and DBOX corpora, and four Dutch-language corpora: the OVIS, DIAMOND, Dutch Map Task[12], and Schiphol [13] corpora. Addition is foreseen of dialogues from the AMI corpus, the YP corpus, and several other corpora.

4.4.1 Map Task and DBOX Dialogues

The Map Task and DBOX dialogues in the DialogBank were annotated using the ANVIL tool in which a facility has been created to export annotations in the DiAML-XML reference format of ISO 24617-2 (see Bunt et al., 2012).

[12] See `http://doc//.ukdataservice.ac.uk/doc/4632/mrdoc/pdf/4632userguide.pdf`.

[13] See Prüst et al., 1984.

Example (14) in the appendix shows the result for a very short dialogue fragment. This format is perfect for machine consumption, but rather inconvenient for human readers, for example for checking the correctness of annotations. The more compact tabular formats shown below are more attractive in that respect.

The DBOX application (quiz game dialogues) called for some small extensions to the ISO annotation scheme, which were made in accordance with the guidelines included in the ISO 24617-2 standard for extending the annotation scheme. Two additional dimensions were introduced: Task Management (also familiar from DAMSL), for dialogue acts where the rules of the game are discussed, and Contact Management, also familiar from DIT^{++}, for dialogue acts where the participants establish, check, or end contact between them.

4.4.2 Switchboard Dialogues

The dialogues in the Switchboard corpus were originally represented in a 3-column tabular format where the leftmost column contains an identifier of the slash unit in the third column, in and the middle column contains a SWBD-DAMSL function tag.[14] In constructing the SWBD-ISO corpus, all in-line markups of filled pauses were replaced by *Stalling* tags and in-line markups of restarts by *SelfCorrection* tags. The result looks as shown in (13).

(13)

sw01-0105-0001-A001-01	setQuestion	A.1 utt1: Jimmy, {D so } how do you get most of your news? /
sw01-0105-0002-B002-01	stalling selfCorrection stalling answer	B.2 utt1: {D Well, } [I kind of, + {F uh, } I] watch the, {F uh, } national news everyday, for one /
sw01-0105-0003-B002-02	answer	B.2 utt2: I also read one or two papers a day /
sw01-0105-0004-B002-03	selfCorrection inform	B.2 utt3: {C and } [I'm a, + I'm pretty much a] news junkie /
sw01-0105-0005-B002-04	answer	B.2 utt4: {C and } I tune in to CNN a lot./
sw01-0105-0006-A003-01	autoPositive	A 3 utt1: {F Oh, } wow /

While convenient for human readers, this format is not optimal for computer processing. The numbering of speaker turns and slash units is redundant (and turns have no special status in the ISO standard), and the rightmost column contains a mixed bag of information types (speaker, turn number, slash unit number within turn, transcribed slash unit, and disfluency and other markups). It could be converted to an XML representation like DiAML-XML by interpreting the first column as the values of the `xml:id` attribute, the second as the values of the `communicativeFunction` attribute,

[14] The Switchboard corpus is also available in NXT format, without in-line markups (see Calhoun et al., 2010.

and the third as the values of the `sender` and `target` attributes and the textual rendering of slash units. However, representations like (13) differ from DiAML-annotations in three fundamental respects: (1) slash units do not always correspond to functional segments, which in general form a more fine-grained way of segmenting a dialogue; (2) the use of in-line markups goes against the ISO requirement that annotations should be in stand-off form; and (3) annotations according to ISO 24617-2 contain more information than just communicative functions, in particular also dimensions, qualifiers, and dependence relations, which are semantically indispensable.

These differences are taken into account in the design of a tabular representation format, called 'DiAML-TabSW', that is relatively close to that of (13), and facilitates comparison between the SWBD-DAMSL and ISO annotation schemes. for incorporating annotated Switchboard dialogues into the DialogBank, first, existing annotated dialogues were re-segmented into functional segments, and the functional segments that do not correspond to a slash unit were newly annotated with ISO 24617-2 communicative function tags and dimension tags. Second, a copy of was made of the slash unit transcriptions in which all in-line markups were interpreted in terms of communicative functions, rhetorical relations, or qualifiers whenever possible, and removed. Third, the functional segments were represented in stand-off fashion by referring to a file that contains segment definitions in terms of word tokens or time points. Finally, the annotations of functional segments were enriched with functional and feedback dependences, qualifiers, and rhetorical relations.

Figure 1 shows the resulting representation. The first four columns represent the annotations proper: (1) functional segment identifiers; (2) dialogue act identifiers; (3) dialogue acts; and (4) sender, with much of the information concentrated in the third column: dimension, communicative function, dependences (as in "Ta:answer (da2)"), qualifiers and rhetorical relations. The fifth and sixth columns, containing funsctional segment texts and turn transcripts, have been added for the convenience of human readers, and have no formal status.

4.4.3 Other Annotated Dialogues and their Representation

The dialogues in the DIAMOND corpus were originally annotated with the DIT^{++} annotation scheme, for which the DitAT annotation tool was developed (Geertzen, 2007); this tool produces representations in a multi-column tabular format with a separate column for each dimension. For inclusion of ISO 24617-2 versions of these annotations in the DialogBank, a new multi-column tabular format was defined, the 'DiAML-MultiTab' format, with one column identifying functional segments in stand-off fashion, as in the DiAML-

markables	ID	Dialogue acts	Sp	FS text	Turn transcript
sw01-0105-fs.1	da1	Ta:setQuestion	A	Jimmy, so how do you get most of your news?	Jimmy, {D so } how do you get most of your news? /
			B		{D Well, } [I kind of, + {F uh, } I] watch the, national news every day, for one / I also read one or two papers a day / {C and } [I'm a, + I'm pretty much a] / news junkie {C and } I tune in to CNN a lot /
sw01-0105-fs.2	da2 da3	TiM:stalling TuM:turnTake	B	Well,	
sw01-0105-fs.3	da4	OCM: selfCorrection	B	I kind of, I	
sw01-0105-fs.4	da5	TiM;stalling	B	uh	
sw01-0105-fs.5	da6	Ta:answer (Fu:da1)	B	I watch the national news every day, for one	
sw01-0105-fs.6	da7	TiM:stalling	B	uh	
sw01-0105-fs.7	da8	Ta:answer (Fu:da1) {Expansion: foregr da7}	B	I also read one or two papers a day	
sw01-0105-fs.8	da9	TuM:turnKeep	B	and	
sw01-0105-fs.9	da10	OCM: selfCorrection	B	I'm a, I'm pretty much a	
sw01-0105-fs.10	da11	Ta:inform	B	I'm pretty much a news junkie	
sw01-0105-fs.11	da12	TuM:turnKeep	B	and	
sw01-0105-fs.12	da13	Ta:answer (Fu:da2) {Expansion: foregr da7, da9}	B	I tune in to CNN a lot	
sw01-0105-fs.13	da14	AuF:autoPositive (Fe: da6 da8 da13)	A	Oh, wow.	Oh, wow

Fig. 1 ISO 24617-2 annotation of dialogue fragment in example (13), represented in DiAML-TabSW format. (Ta = Task, TiM = Time Management, TuM = Turn Management, OCM = Own Communication Management, AuF = Auto-Feedback)

TabSW format, one column indicating the speaker, and one column per dimension for representing communicative functions, qualifiers, dependence relations and rhetorical relations. Figure 2 illustrates this format, which was proven to be convertible without loss of information to DiAML-XML and vice-versa (Bunt et al., 2016). In the example, those columns have been suppressed that correspond to dimensions in which no communicative functions were marked up for this fragment.

The DiAML-MultiTab format was used also for representing re-annotated dialogues from the OVIS and TRAINS corpora, and newly annotated Schiphol dialogues.

5 Conclusions and Perspectives

The ISO 24617-2 standard for dialogue annotation has as its main features a rich taxonomy of clearly defined communicative functions, including many functions from previously developed annotation schemes such as DAMSL, DIT^{++}, and ICSI-MRDA; the distinction of nine dimensions, inherited from the DIT^{++} schema; functional and feedback dependence relations that account for semantic dependences between dialogue acts; the use of qualifiers for expressing (un-)certainty, conditionality and sentiment; and rhetorical relations among dialogue acts. In this chapter. experiences and experiments were discussed that investigate how these features play out in human and automatic dialogue annotation.

mark-ables	sp	fs text	turn transcript	Task	Auto-Feedback	Turn Man.	Time Man.	Discourse Struct.	SocialObl. Man.
			hello, can I help you						
TR1-fs.1	s	hello							da1:Init. Greeting
TR1-fs.2	s	can I help you						da2:Offer	
			uhm, yes hello,maybe, I'd like to take a tanker...						
TR1-fs.3	u	uhm				da3: Turn Take	da4: Stalling		
TR1-fs.4	u	yes hello			da5:Pos. (Fe:da1)				
TR1-fs.5	u	yes maybe						da6: Accept Offer [uncertain] (Fu:da2)	
TR1-fs.6	u	I'like to take...		da7: Inform					

Fig. 2 ISO 24617-2 annotation of TRAINS dialogue fragment represented in DiAML-MultiTab format

New and emerging corpora were discussed that contain dialogues, annotated according to the ISO 24617-2 standard, notably the DBOX, YP, and DialogBank corpora. Such resources offer a promising basis for the study of human communication as well as for the design and training of modules in dialogue systems, such as recognizers of communicative functions in human interactive behaviour, and dialogue managers in speech-based or multimodal dialogue systems.

References

Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, and B. S. . M. Siegel (1998). *Dialogue acts in VERBMOBIL-2 (second edition). Verbmobil Report 226.* Saarbrücken: DFKI.

Allen, J. and M. Core (1997). *DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). Technical Report.* Rochester, NY: University of Rochester.

Allwood, J. (1992). *On dialogue cohesion.* Gothenburg University, Department of Linguistics.

Bunt, H. (1994). Context and dialogue control. *Think Quarterly 3*(1), 19–31.

Bunt, H. (2000). Dialogue pragmatics and context specification. In H. Bunt and W. Black (Eds.), *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics.*, pp. 81–150. Amsterdam: John Benjamins.

Bunt, H. (2006). Dimensions in dialogue annotation. In *Proceedings 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy*, Paris. ELRA.

Bunt, H. (2009). The DIT^{++} taxonomy for functional for dialogue markup. In D. Heylen, C. Pelachaud, R. Catizone, and D. Traum (Eds.), *Proceedings of EDAML-AAMAS Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts"*, Budapest, pp. 36–48.

Bunt, H. (2011). Multifunctionality in dialogue. *Computer, Speech and Language 25*, 222–245.

Bunt, H. (2015). On the Principles of Semantic Annotation. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, London (ISA-11).*, pp. 1–13.

Bunt, H., J. Alexandersson, J. Carletta, J.-W. Choe, A. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum (2010). Towards and ISO standard for dialogue act annotation. In *Proceedings 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta. Paris. ELDA.

Bunt, H., J. Alexandersson, J.-W. Choe, A. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings 8th International*

Conference on Language Resources and Evaluation (LREC 2012), Istanbul. ELDA.

Bunt, H., A. Fang, J. Cao, X. Liu, and V. Petukhova (2013). Issues in the addition of ISO standard annotations to the Switchboard corpus. In *Proceedings 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, Potsdam, pp. 59–70.

Bunt, H., M. Kipp, and V. Petukhova (2012). Using DiAML and ANVIL for multimodal dialogue annotation. In *Proceedings 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul. ELRA, Paris.

Bunt, H., V. Petukhova, A. Malchanau, and K. Wijnhoven (2016). The DialogBank. In *Proceedings 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia. Paris: ELRA.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver (2010). The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation 44(4)*, 387–419.

Carletta, J., S. Isard, J. Kowtko, and G. Doherty-Sneddon (1996). *HCRC dialogue structure coding manual.* University of Edinburgh. Technical Report HCRC/TR-82.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin 70*, 213–261.

Dhillon, R., S. Bhagat, H. Carvey, and E. Schriberg (2004). *Meeting recorder project: dialogue labelling guide. ICSI Technical Report TR-04-002.* University of California at Berkeley.

Fang, A., J. Cao, H. Bunt, and X. Liu (2011). Relating the semantics of dialogue acts to linguisitic properties: A machine learning perspective through lexical cues. In *Proceedings IEEE-ICSC 2011 Workshop on Semantic Annotation for Computational Linguistic Resources*, Stanford, CA.

Fang, A., J. Cao, H. Bunt, and X. Liu (2012a). The annotation of the Switchboard corpus with the new ISO standard for dialogue act analysis. In *Proceedings 8th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-8)*, Pisa, pp. 13–18.

Fang, A., J. Cao, H. Bunt, and X. Liu (2012b). Applicability Verification of a New ISO Standard for Dialogue Act Annotation with the Switchboard Corpus. In *Proceedings of EACL 2012 Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, Avignon.

Geertzen, J. (2007). DitAT: A flexible tool to support web-based dialogue annotation. In *Proceedings 7th International Conference on Computational Semantics (IWCS-7)*, Tilburg, pp. 320–323.

Geertzen, J. and H. Bunt (2006). Measuring annotator agreement in a complex, hierarchical dialogue act schema. In *Proceedings SIGDIAL 2006*, Sydney.

Geertzen, J., Y. Girard, and R. Morante (2004). The DIAMOND project. In *Proc. 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004)*, Barcelona.

Geertzen, J., V. Petukhova, and H. Bunt (2008). Evaluating dialogue act tagging with naive and expert annotators. In *Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech. Paris. ELDA.

Hovy, E. and E. Maier (1995). *Parsimonious or profligate: how many and which discourse structure relations? ISI research report.* Marina del Rey: Information Sciences Institute, University of Southern California.

Ide, N. and L. Romary (2004). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering 10*, 211–225.

ISO (2011). *ISO 24612: Language Resource Management - Linguistic Annotation Framework (LAF).* Geneva: ISO.

ISO (2012). *ISO 24617-2: Language Resource Management - Semantic Annotation Framework (SemAF) - Part 2: Dialogue Acts.* Geneva: ISO.

ISO (2015). *ISO 24617-6: Language Resource Management - Semantic Annotation Framework (SemAF) - Part 6: Principles of Semantic Annotation.* Geneva: ISO.

Jurafsky, D., E. Shriberg, and D. Biasca (1997). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation: Coders Manual, Draft 1.3.* University of Colorado.

Lendvai, P., A. van den Bosch, E. Krahmer, and S. Canisius (2004). Memory-based robust interpretation of recognised speech. In *Proceeings 9th International Conference on Speech and Computer (SPECOM'04)*, St. Petersburg, pp. 415–422.

Lesch, S., T. Kleinbauer, and J. Alexandersson (2005). A new metric for the evaluation of dialog act classification. In *Proceeings 9th Workshop on the Semantics and Pragmatics of Dialogue (DIALOR)*, Nancy.

Meteer, M. and A. Taylor (1995). *Dysflency Annotation Stylebook for the Switchboard Corpus.* Washington: Linguistic Data Consortium.

Petukhova, V. (2011). *Multidimensional Dialogue Modelling. Ph.D. Dissertation.* Tilburg University.

Petukhova, V. and H. Bunt (2011). Incremental dialogue act understanding. In *Proceedings Ninth International Conference on Computational Semantics (IWCS 2011), Oxford*, pp. 235 – 244.

Petukhova, V., M. Gropp, D. Klakow, G. Eigner, M. Topf, S. Srb, P. Moticek, B. Potard, J. Dines, O. Deroo, R. Egeter, U. Meinz, S. Liersch, and A. Schmidt (2014). The DBOX corpus collection of spoken human-human and human-machine dialogues. In *Proceedings 9^{th} International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Petukhova, V., L. Prévot, and H. Bunt (2011). Multi-level discourse relations between dialogue units. In *Proceedings 6th Joint ACL-ISO workshop on Interoperable Semantic Annotation (ISA-6)*, Oxford, pp. 18–28.

Popescu-Belis, A. (2005). *Dialogue Acts: One or More Dimensions? ISSCO Working Paper 62*. Geneva: ISSCO.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008). The Penn Discourse TreeBank 2.0. In *Proceedings 6th International Conference on Language Resources and Systems (LREC 2008)*, Marrakech.

Prüst, H., G. Minnen, and R.-J. Beun (1984). *Transcriptie dialooogesperiment juni/juli 1984, IPO Rapport 481*. Institute for Perception Research, Eindhoven University of Technology.

Reithinger, N. and M. Klesen (1997). Dialogue act classificationusing language models. In *Proceedings of Eurospeech-97*, pp. 2235–2238.

Samuel, K., S. Carberry, and K.Vijay-Shanker (1998). Dialogue act tagging with transformation-based learning. In *Proceedings ACL 1998*, Montreal, pp. 1150–1156.

Stolcke, A., K. Res, K. Coccaro, E. Shriberg, R.Bates, D.Jurafsky, P.Taylor, R.Martin, C. van Ess-Dykema, and M. Meteer (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Lingsuitics 26(3)*, 339–373.

Traum, D. (2000). 20 questions on dialogue act taxonomies. *Journal of Semantics 17*(1), 7–30.

Zimmermann, M., Y. Lui, E. Shriberg, and A. Stolcke (2005). Toward joint segmentation and classification of dialogue acts in multiparty meetings. In *Proceedings of the Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI-05)*, Berlin, pp. 187–193. Springer.

Appendix

This appendix shows the ISO 24617-2 annotation of the first two utterances of a Map Task dialogue in the DialogBank corpus, as produced with the ANVIL tool and exported in DiAML format. In a TEI-compliant way,[15] the first part identifies the two dialogue participants ("p1" and "p2"), followed by a second part that identifies the word tokens in the audio-video input stream, and a third part that identifies the functional segments in terms of the word tokens. The last part represents the dialogue act annotations in the DIAML format of the ISO standard.

(14) G: right
G: go south and you'll pass some cliffs on your right

[15] Text Encoding Initiative: www.tei.org

F: okay

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <profileDescr xmlns="">
    <particDescr xml:id="p1">
      <p>the 1. participant</p>
    </particDescr>
    <particDescr xml:id="p2">
      <p>the 2. participant</p>
    </particDescr>
  </profileDescr>
  <text>
    <body />
    <div>
      <head>The dialogue turns, segmented into words
(TEI-compliant)</head>
      <u>
      <w xml:id="w1">right</w>
        <w xml:id="w2">go</w>
        <w xml:id="w3">south</w>
        <w xml:id="w4">and</w>
        <w xml:id="w5">you'll</w>
        <w xml:id="w6">pass</w>
        <w xml:id="w7">some</w>
        <w xml:id="w8">cliffs</w>
        <w xml:id="w9">on</w>
        <w xml:id="w10">your</w>
        <w xml:id="w11">right</w>
        <w xml:id="w12">okay</w>
        ...
        </u>
      </div>
    <div>
      <head>Identification of functional segments</head>
      <spanGrp xml:id="ves1" type="functionalVerbalSegment">
        <span xml:id="ts1" type="textStretch" from="w1" to="w1" />
      </spanGrp>
      <fs type="functionalSegment" xml:id="fs1">
        <f name="verbalComponent" fVal="#ves1" />
      </fs>
      <spanGrp xml:id="ves2" type="functionalVerbalSegment">
        <span xml:id="ts2" type="textStretch" from="w2" to="w11" />
      </spanGrp>
     <fs type="functionalSegment" xml:id="fs2">
        <f name="verbalComponent" fVal="#ves2" />
      </fs>
      <spanGrp xml:id="ves3" type="functionalVerbalSegment">
        <span xml:id="ts3" type="textStretch" from="w12" to="w12" />
      </spanGrp>
     <fs type="functionalSegment" xml:id="fs3">
        <f name="verbalComponent" fVal="#ves3" />
      </fs>
    </div>
```

```
    <diaml xmlns="http://www.iso.org/diaml">

    <dialogueAct xml:id="da1"
    target="#fs1" sender="#p1" addressee="#p2"
    dimension="turnManagement" communicativeFunction="turnTake"  />

    <dialogueAct xml:id="da2"
    target="#fs1" sender="#p1" addressee="#p2"
    dimension="discourseStructuring" communicativeFunction="opening"  />

    <dialogueAct xml:id="da3"
    target="#fs2" sender="#p1" addressee="#p2"
    dimension="task" communicativeFunction="instruct"  />

    <dialogueAct xml:id="da4"
    target="#fs3" sender="#p2" addressee="#p1"
    dimension="autoFeedback" communicativeFunction="autoPositive"
    feedbackDependence="#fs2" />
    </diaml>

 </text>
</TEI>
```