The coding and annotation of multimodal dialogue acts

Volha Petukhova* and Harry Bunt**

*Human Speech and Language Technologies, Vicomtech-IK4, Spain
** Tilburg Center for Cognition and Communication, Tilburg University, The Netherlands
vpetukhova@vicomtech.org; harry.bunt@uvt.nl

Abstract

This paper describes how the ISO 24617-2 annotation scheme can be used, together with the DIT^{++} method of 'multidimensional segmentation', to annotate nonverbal and multimodal dialogue behaviour. We analyse the fundamental distinction between (a) the coding of surface features; (b) form-related semantic classification; and (c) semantic annotation in terms of dialogue acts, supported by experimental studies of (a) and (b). We discuss examples of specification languages for representing the results of each of these activities, showing how dialogue act annotations can be attached to XML representations of functional segments of multimodal data.

Keywords: multimodal dialogue behaviour, multimodal dialogue act annotation, international standards

1. Introduction

Recent years have witnessed a growing interest in annotating linguistic data at the semantic level, including the annotation of dialogue corpus data. Various annotation schemes have been developed for dialogue act annotation, of which DAMSL (Dialogue Act Markup using Several Layers, Allen and Core (1997)) is perhaps the most widely used. The DIT^{++} scheme (Bunt, 2006; 2009) combines the multidimensional DIT scheme (Bunt, 1994) with concepts from DAMSL and various other schemes, providing precise definitions for its communicative functions and dimensions. This scheme has been the basis for defining the international standard for dialogue act annotation ISO 24617-2.[1]. This annotation scheme is designed in a such way that it can be applied not only to spoken dialogue, as is the case for most of the previously defined dialogue annotation schemes, but also to multimodal dialogue, as shown in this paper.

Before going into the details of multimodal dialogue act annotation, we first discuss the fundamentals of *'coding'* (or 'transcribing') and *'annotating'* multimodal dialogue behaviour, making a clear distinction between the two (Section 2). We subsequently discuss practical aspects of transcription, segmentation and annotation processes. We report on a series of coding experiments, performed in order to measure the reliability of human codings, and compare these with automatic coding reliabilities reported in the literature. Section 3 describes the coding of multimodal dialogue behaviour in terms of low-level surface features of body movements. Section 4 addresses the form-related classification of visible movements by humans and machines. Section 5 discusses the multidimensional segmentation and annotation of multimodal data using the ISO 24617-2 dialogue act annotation scheme. Section 6 presents the XML-based representation of multimodal dialogue data, using the Dialogue Act Markup Language (DiAML) defined in ISO 24617-2. Section 7 concludes the paper with a brief look back and a look forward to future work in this area.

[1]ISO Draft International Standard DIS 24617-2:2010 has been accepted as an ISO standard in January 2011.

2. Coding communicative behaviour

Accurate dialogue act annotation requires precise transcription of communicative behaviour in dialogue. Transcriptions of spoken dialogue are either obtained through automatic speech recognition, possibly with manual correction, or performed manually by trained transcribers. In many dialogue corpora, such as **MapTask**[2], **AMI**[3], and **TRAINS**[4], speech transcriptions are provided in orthographical form with word-level timings.

Prosodic properties of dialogue contributions can be computed automatically using tools for voice analysis, of which PRAAT[5] is perhaps the most widely used. Properties that can be computed by PRAAT include minimum, maximum, mean, and standard deviation of *pitch* (F0 in Hz), *energy* (RMS), *voicing* (fraction of locally unvoiced frames and number of voice breaks) and *speaking rate* (number of syllables per second). Both raw and normalized versions of these features may be used. Speaker-normalized features can be obtained by computing z-scores (z = (X-mean)/standard deviation), where mean and standard deviation are calculated from all functional segments produced by the same speaker. Normalizations by first speaker turn and by prior speaker turn are also used. Additionally, temporal and durational properties can be calculated: token *duration* and *floor-transfer offset*[6], computed in milliseconds. Prosodic transcriptions may contain manually identified and labeled tonal events using coding systems like ToBI (Beckman et al., 2005).

For the coding of gestures and other movements various schemes have been designed which support the characterization of movements in terms of low-level (or surface)

[2]Detailed information about the MapTask project can be found at http://www.hcrc.ed.ac.uk/maptask/

[3]Augmented Multi-party Interaction, for more information visit http://www.amiproject.org/

[4]For more information about the TRAINS corpus please visit http://www.cs.rochester.edu/research/speech/trains.html

[5]For more information and downloads visit http://praat.org

[6]Difference between the time that a turn starts and the moment that the previous turn ends.

behavioural features, such as changes in muscular activity (body parts involved and form of movement), or direction, trajectory and speed of movements. For example, the Facial Action Coding System (FACS)[7] codes facial expressions describing muscular activities that produce changes in facial appearance. HamNoSys[8] is a transcription system for coding hand gestures by describing shape, direction, speed, length and form of movement, hands orientation and location. The CoGest scheme (Gut et al., 2003) proposes a feature-based vector notation where a gesture is represented by a set of values of gesture attributes like source, trajectory, target, and shape of trajectory. For example, the CoGest string $15m, 5A, ri, ci, 1B, l, r(0), me, 15m, 5A, rp$ describes an unrepeated gesture carried out with medium speed with the right hand tracing a large circle with a pointed index finger, which starts and ends with the hands on the lap.

Other schemes use types of movement as labels, e.g. nod or shake for the head (e.g. MUMIN, Allwood et al., 2004) or iconic, metaphoric, adaptor for the hand (e.g. Kipp, 2004). Still others immediately assign a high-level semantic interpretation to observed movements, e.g. concordance signal, negative signal, turn signal (see the e.g. AMI Guidelines for Individual Actions Annotation, 2005).

The transcription of the movements of dialogue participants in terms of pragmatic meaning is risky for several reasons. First of all, the meanings that different movements may convey should be established empirically, rather than *a priori*. There are too many different possible body movements, facial expressions, and gestures, with considerable cultural and even individual variation to be able to judge their meaning unequivocally as part of a transcription scheme. The characterisation of movements in terms of surface features, by contrast, allows their interpretation to be tested empirically, determining e.g. which behaviour means disbelief, agreement or puzzlement. In other words, *description* should be separated from *interpretation*. Similar considerations apply to the automatic generation of communicative behaviour. When behaviour with certain functional meaning is to be produced, we need to know what behavioural features correspond to this type of behaviour.

Second, manual coding is expensive. Automatic speech recognizers can be used for the production of speech transcriptions (with manual correction of the output). Automatic detection and coding of visible movements either from video or from direct observation is an active research field with applications in a range of domains such as virtual reality, 'smart' surveillance systems, advanced user interfaces, motion analysis, and robotics. The state of the technology for markerless motion capture is mature enough to boost the research on the recognition of action units using off-the-shelf and affordable equipment, such as webcams for facial expression tracking (Dornaika and Davoine, 2006; Dornaika and Raducanu, 2009), and depth sensing devices for full-body tracking (Shotton et al., 2009). In this area, description and interpretation are distinguished clearly. The standard procedure goes as follows:

1. sensors capture any noticeable motions, and low-level features are derived and selected, such as parts of body moved, relative and absolute positions, angles, velocity, periodicity, and intensity.Lösch (2006) extracted for example 320 features from the body model which is provided by the tracking system;
2. potential action spotting: segmentation of the data stream into temporal regions that might correspond to actions or transitions from one action to another (see e.g. Stiefmeier and Roggen, 2007);
3. semantic interpretation: classification of action units, associating motion segments with categories of a knowledge base, e.g. pointing gesture, smile, leaning forward, kissing, or shaking hands, or identifying an unknown motion pattern through classification procedures such as HMMs (Ahmad and Lee, 2006), DTW (Kang et al., 2006), or SVM (Ramanan and Forsyth, 2003);
4. pragmatic interpretation, e.g. in terms of communicative functions such as agreement, or in terms of communicative function qualifiers (see Section 4) such as uncertainty, anger, happiness, surprise, or fear (De la Torre and Cohn, 2011).

The most recent approaches tend to merge the stages of segmentation and classification, i.e. to segment while classifying (see Unzueta and Goenetxea, 2010; Zhou et al., 2011). This speeds up the action recognition process as a whole. On such an approach there is no segmentation without the classification of units, since the identification of action unit boundaries depends on how an action unit is defined.

While the combination of segmentation and annotation has practical advantages, the distinction between *description* or *coding*[9] and *annotation* is methodologically very important. Coding is the representation of speech, sound, or movement using a certain coding system, e.g. phonetic or orthographic transcriptions for speech, and representation for physical realization of body and facial actions. For the latter, a variety of markup languages have been created, such as the Virtual Human Markup Language (VHML)[10] and the Multimodal Utterance Markup Language (MURML, Kranstedt et al., 2002). The Behaviour Markup Language (BML) developed within the SAIBA[11]

[7] For more information visit: `http://face-and-emotion.com/dataface/general/homepage.jsp`

[8] For more information visit `http://www.sign-lang.uni-hamburg.de/projekte/hamnosys/hamnosyserklaerungen/englisch/contents.html`

[9] Both terms are used in the literature. We prefer to use the term 'coding', since this term in our view better captures the essence of this process, namely, representation of perceived bodily actions using a specific notation system, e.g. feature vectors. Coding results in transcription.

[10] See `http://www.vhml.org`

[11] The Situation, Agent, Intention, Behavior, Animation framework specifies multimodal generation at a macro-scale, consisting of processing stages on three different levels: (1) planning of a communicative intent, (2) planning of a multimodal realization of this intent, and (3) realization of the planned behaviour. For more information please visit `http://www.mindmakers.org/projects/SAIBA`

framework is a description language for controlling the verbal and nonverbal behaviour of virtual characters (see Kopp et al., 2006 and Vilhjalmsson et al., 2007). It describes the physical realization of behaviours and synchronization constraints. When extended properly with articulate specifications of the surface form of nonverbal behaviour, BML can also be used for coding human multimodal dialogue behaviour.

The term 'annotation' refers to the addition of linguistic information to segments of language data and/or nonverbal communicative behaviour (see the ISO Linguistic Annotation Framework, ISO 24612:2010), where linguistic information may be (morpho-)syntactic, semantic or pragmatic. As part of the ISO 24617-2 standard, the Dialogue Act Markup language (DiAML) for dialogue act annotation has been defined, which will be discussed in Section 6.

3. Obtaining reliable multimodal transcriptions: coding experiments

When we agree on the importance of the distinction between coding and annotation, and that the former should be performed in terms of behavioural surface features, the question arises what surface feature can be coded reliably. The coding of a movement normally consists of determining (i) body parts involved in the movement; (ii) temporal boundaries and duration of phases of a movement, where often three or four phases are considered: (1) the onset or preparation; (2) the peak, sometimes divided in two (a) stroke, (b) hold), and (3) offset (or retraction); (iii) spatial characteristics like angles, direction, trajectory, distance from and relative position to the rest of the body or specific other body parts, and size; and (iv) characteristics like velocity, periodicity and intensity.

Transcriptions of multimodal communicative behaviour are mostly obtained by employing trained transcribers. This method is expensive and as we will show not all features can be coded reliably by human transcribers. The CoGest scheme provides an elaborate coding system that includes coding of all the surface features for hand and arm gestures mentioned above. Gut et al. (2003) reported that the observed agreement on hand and arm gesture classification when applying the CoGest scheme was only 23.4%. The main source of disagreement was formed by categories like gesture boundaries, trajectory, size, speed and periodicity of movements. De la Torre et al. (2011) also noticed that average manual error compared to automatic temporal segmentation was within 10-12 frames for the movement offset, and 2 frames for the movement peak when coding facial expressions using FACS.

Jovanovic (2007) reported that coding the focus of attention as derived from head, gaze and posture observations can be done with a very high level of agreement and with very high precision: changes are marked in the *middle* of eye movements between old and new target with α agreement (Krippendorf, 1995) between annotators ranging from 0.84 to 0.95. In order to assess the difficulties and possibilities in coding surface features reliably, we performed coding and evaluation experiments focusing on five forms of nonverbal expression: gaze direction, head movements, hand and arm gestures, posture shifts, and facial expressions.

Two scenario-based dialogues with a total duration of 51 minutes from the AMI corpus[12] were selected. Orthographic transcriptions of the speech were produced semi-automatically (manually corrected output from an automatic speech recognizer). Transcriptions of the movements of each participant were performed fully manually. Transcribers were asked to segment the behaviour (assigning start and end times), and to code surface features such as what *body part* is involved in the action (head, hand, arm, upper body, lips, eyes, eyebrows, chin, nose, etc.), *direction* of movement (up, down, left, right, backward, forward); *trajectory* (e.g. line, circle, arch); *distance* from the body for hands (e.g. close to the body, in contact with the body); *size* (e.g. large, small, medium, extra large); *velocity* (slow, medium, fast); and *periodicity* (number of repetitions up to 20 times). For each movement *intensity* was determined: 0 - no movement; 1 trace (noticeable movement); 2 marked (significant evidence for a movement). The *floor transfer offset* (see footnote 6) and *duration* of a movement (in milliseconds) were computed automatically. The coding was thus in the line with the CoGest scheme.

The nonverbal behaviour of the dialogue participants was transcribed using video recordings for each individual participant, running them without sound to eliminate the influence of what was said. Transcriptions were performed using the ANVIL tool[13], which allows transcriptions in separate tiers for each participant, using specific tiers for each type of movement (see Bunt, Kipp and Petukhova, 2012).

Movements were transcribed by two coders in order to be able to judge the reliability of the coding. Inter-coder agreement was measured in terms of Cohen's kappa. The major disagreements observed between coders concern (1) the definition of temporal boundaries (segmentation); (2) judgements of the velocity and intensity of movements; (3) determination of spatial characteristics such as size, trajectory and distance.

As for temporal segmentation, the difference between annotators ranged between 120 ms (up to $\pm$ 3 frames, e.g. for gaze re-direction) to 520 ms (up to $\pm$ 13 frames, for hand gestures because some neighboring locations may be quite subtle). In terms of kappa, the agreement reached here was moderate: .46. This is comparable with findings reported by De la Torre et al. (2011), discussed above. As for the velocity and intensity of movements, coders have genuine difficulty to judge these rather subjective and speaker-dependent characteristics when no or limited information about the dialogue participants is available. Coding does speed up and judgments are made with higher degree of certainty in the course of the coding process. Agreement between coders in terms of kappa for defining the speed of movements was .29, with differences per expression type: the highest when judging the speed of head nods (.49) and the lowest when judging the speed of facial activities, such as eyebrow or lip movements and blinking (.18). Finally, coders differed in opinion about movement intensity; in particular judgments about 0 (no movement) or 1 (noticeable movement) categories were often dissimilar, one an-

[12] See `http://www.amiproject.org/`

[13] See `http://www.dfki.de/~kipp/anvil`

notator thinking that there was some trace of a movement, another not seeing any movement at all. Overall kappa was .41.

As for spatial features, especially the size of movements is a rather subjective category, and a source of disagreement (kappa .38), with the lowest score for head movements (kappa .11) and the highest for hand and arm gestures (.57). Trajectory labeling caused some confusion (e.g. one coder sees an ellipsis, another a circle or arch), ranging from .36 for head to .21 for arm gestures and .09 for gaze direction. Judging distances, coders have less difficulty (kappa of .53), maybe because the participant's body forms a clearer reference point.

Spatial characteristics of body movements are very important for their interpretation, since the same type of movement performed with different speed, amplitude or periodicity may have different communicative functions (see e.g. Petukhova and Bunt, 2010b). Temporal features are obviously of crucial importance for the synchronization of verbal and nonverbal behaviour, in particular when this is used for the generation of multimodal dialogue utterances.

While human coding is seen not to be reliable, automatic techniques, by contrast, are quite robust, offering optimal metrics to segment a video stream into action units (see e.g. De la Torre et al., 2011), to measure the speed, size and intensity of image change, and to calculate the trajectory and distance of movements (see, e.g. Lösch et al., 2008 and Zhou et al., 2011). Moreover, automatic techniques provide statistical features dealing with variations of position, distance, velocity and intensity relative to the body and to extrinsic objects. This is a good news, and gives some hope for the reliable recognition of these features in the future.

The main lesson to be learned here is that humans are generally not very successful in coding spatiotemporal characteristics of body movements reliably; machines are better at this task and can take this job over in the near future.

4. Form-related interpretation of visible movement: annotation experiments

Movements, transcribed as discussed in the previous section, can be assigned a meaning in terms *type* of movement. For example, an up and down head movement is a nod, a left to right head movement is a head shake, and elongating the lips and lifting the lip corners is a smile. Annotation of the type of transcribed movements allows the determination of variations (such as different spatial, temporal, durational and intensity qualities) in bodily activity that may have one and the same meaning. This information provides an empirical basis for precise semantic/pragmatic analysis, e.g. to establish whether one and the same type of movement but with different low-level characteristics may have the same or a slightly different communicative function, which is equally important for the interpretation of dialogue behaviour and for its generation.

To assess the reliability of human determination of type of visible movement, we performed an experiment for which a classification scheme was designed that combines the MUMIN scheme (Allwood et al., 2004) and the scheme provided with the ANVIL tool (Kipp, 2004), and makes some extensions. We defined 84 movement types: 2 for gaze, 9 for head movements, 40 for hand and arm movements, 24 for facial expressions and 9 for posture shifts. Coders were asked to also indicate their degree of certainty for each decision that they made, ranging from 0 (not certain at all) to 5 (very certain).

Table 1: Cohen's kappa scores for each type of visible movement reached by two coders.

Type of expression	Kappa
Gaze	.83
Head movements	.82
Hand movements	.48
Facial expression	.65
Posture shifts	.81

The experiments show that humans are good at action classification (see Table 1) and are quite certain in making such decisions (3.8 average degree of certainty). As a rule they do not experience any problems in identifying movement types.

Table 1 shows that the classification of arm and hand movements is a relatively difficult task. A major source of disagreement here was the classification of hand shapes, e.g. what one annotator sees as a open palm gesture with all fingers in joined position and bended, another sees all fingers joined except for the thumb, but not bended.

People have a richer experience and background knowledge for action classification than machines. Machines cannot operate directly in terms of form-related classes, but when provided with a sufficiently large variety of examples of one and the same type of movement, machines can learn this, as shown by Lösch et al., (2008) for teaching robots to perform certain types of action. When recognizing actions, the machine task is often just to identify similar surface patterns and mark them; the marked patters are then classified by experts, and this information is fed back into the system for the next recognition iteration, this time in terms of action types (see e.g. Zhou et al., 2011).

The main conclusion from these experiments is that surface features of nonverbal behaviour can be interpreted reliably by human transcribers in terms of type of visible movement. Machines can use such annotations to learn to interpret movement features. Together with motion tracking features, which can be computed automatically with high precision, such annotations are useful for identifying and annotating the meaningful units in dialogue in terms of dialogue acts, resulting in more accurate and adequate analysis of dialogue behaviour, as we will discuss in the next two sections.

5. Segmentation and annotation of multimodal dialogue acts

Communication in multimodal dialogue is a complex activity. Figure 1 shows that dialogue participants most of the time perform some communicative activity. By re-directing his gaze from the working table to participant D, who is speaking, and shifting his posture to working position, participant B indicates that he is paying attention; by a short single head nod and lip movements he signals that he un-

Speaker		Observed communicative behaviour/ annotation								
coding	D	words	What's	teletext						
		gaze	averted(table)	person B						
		eyes		narrow						
		posture	working position							
annotation		Feedback	SetQuestion							
		TurnM.		Turn assign to B						
coding	B	words				um	It's	a	British	thing
		gaze	averted(table)	person D						
		head		short single nod						
		lips		random movements						
		posture	down	working position						
annotation		Feedback		pos. attention						
		TurnM.		turn accept	turn keep					

Figure 1: Example of coding and annotating multimodal dialogue behaviour.

Speaker			Observe communicative behaviour			
A	speech		He	kissed	me	
	gaze		averted	direct to B		
	face	forehead		relaxed		
		eyebrows		Half-raised		
		eyes		narrowed, corners wrinkled		
		cheeks		outer, upper area of cheeks raised		
		lips		elongated, both corners up		
		chin		No noticeable movement		
	Hand /arm	part		Lower-arm, hand		
		handness		right		
		Hand shape		pointing index finger		
		direction		up		
		trajectory shape		arch		
		location		cheek, right		
		velocity		medium		
		size		large		
		intensity		significant		
annotation	Task		Inform (sem.content: He kissed me *on my right cheek*) Happy			

Figure 2: Example of annotation of multimodal dialogue behaviour.

derstood that D wants B to be the next speaker (D looks at B while asking a question) and accepts the turn.

Nonverbal behaviour may serve several purposes. It may emphasize or articulate the semantic content of a spoken dialogue act as shown in Figure 2 where the pointing gesture to the right cheek contributes to the semantic content of the verbal utterance *He kissed me*, specifying that the kiss was *on the right cheek*.

Nonverbal behaviour may emphasize or support the intended meaning of synchronous verbal behaviour. In the same example in Figure 2 the fact that the speaker was smiling indicates that he liked being kissed: *He kissed me on the right cheek and I liked it*.

Nonverbal behaviour may also perform separate dialogue acts in parallel to what is contributed by another participant. For instance, the majority of head nods signal positive feedback; gaze aversion often signals hesitation and turn keeping (see Figure 1).

Finally, nonverbal behaviour may express a separate dialogue act in parallel to what the same speaker is expressing verbally, adding to the multifunctionality of dialogue utterances. For instance, speech-focused movements accompanying content words (e.g. iconic gestures accompanying the search for a word), or body-focused movements like rubbing cheeks when searching for an elusive word, indicate that the speaker needs some time to gather his/her thoughts or to formulate an utterance, and is therefore stalling for time, while keeping the turn (see e.g. Petukhova and Bunt, 2010b).

All this has consequences for segmenting dialogue behaviour into units and assigning meaning to them. Where a functional segment in speech-only dialogue is a stretch of speech, in multimodal dialogue it is a complex structure, made up of stretches of communicative behaviour in each of the modalities that are used. Figure 1 illustrates this: participant D asks a question for clarification while directing his gaze to participant A (at whom he directs the question) and narrowing his eyes as visual support for conveying the intention to get something clarified. The multimodal functional segment in this case consists of the verbal segment *"What's teletext"*, the stretch of gaze behaviour where D redirects his gaze to A, and the stretch of facial expression behaviour where he narrows his eyes.

An attractive solution for how to identify meaningful multimodal dialogue units and specify their meaning accurately has been proposed in ISO standard 24617-2, based on the DIT multidimensional approach to segmentation and annotation of dialogue acts (see Geertzen et al., 2007). ISO 24617-2 defines a dialogue act as

(1) *communicative activity of a participant in dialogue, interpreted as having a certain communicative function and semantic content.*[14]

A communicative function specifies the way semantic content is to be used by the addressee to update his context

[14] A note, added to the definition, remarks that "A dialogue act may additionally have certain functional dependence relations, rhetorical relations, and feedback dependence relations".

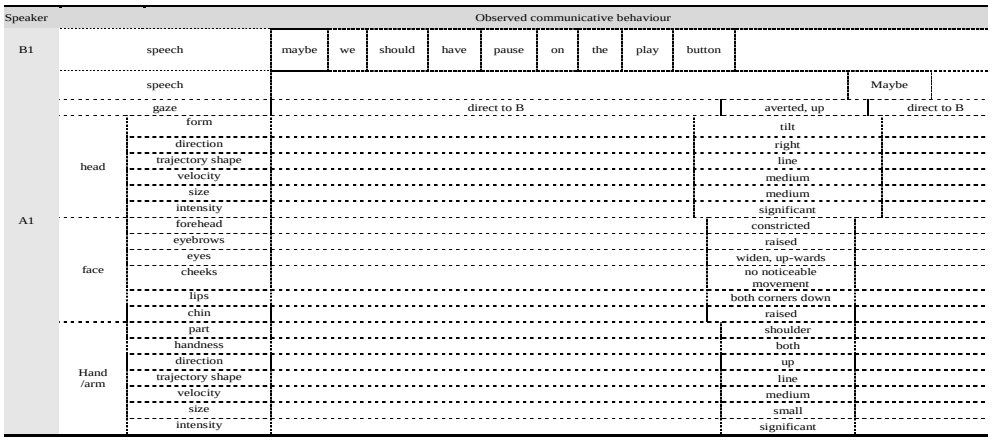

Figure 3: Example of coding multimodal dialogue behaviour.

(2)
```
<timeline unit="ms"
    <when xml:id="t1" absolute="297722"/>
    <when xml:id="t2" absolute="486737"/>
    ...
    <when xml:id="t9" absolute="1897215"/></timeline>
  <head>Verbal contributions, segmented into tokens (TEI-compliant)</head>
  <u>
    <w xml:id="w1">Maybe</w>
    <w xml:id="w2">we</w>
    ...
    <w xml:id="w9">button</w> </u>
  <u>  <w xml:id="w10">Maybe</w></u>
  <fs type="verbalContrib" xml:id="#vec1" vSpan="#ves1" who="#p1 start="#t1" end="#t5"/>
  <spanGrp xml:id="ves1" type="verbalSegment"
   <span xml:id="ts1" type="textStretch" from="#w1" to="#w10"/></spanGrp>
  <fs type="verbalContrib" xml:id="#vec2" vSpan="#ves2" who="#p2 start="#t6" end="#t8"/>
  <spanGrp xml:id="ves2" type="verbalSegment"
   <span xml:id="ts2" type="textStretch" from="#w10" to="#w11"/></spanGrp>
  <kinesic type="gazeBehavr" xml:id="g1" who="#p2" start="#t4" end="#t6" ana="#gad1"/>
  <fs gazeDescr xml:id="gad1">
   <f name="source" fVal="#table"/> <f name="goal" fVal="#p1"/>
   <f name="direction"><symbol value="up-right"/></f></fs>
  <kinesic type="lipMove" xml:id="lip3" who="#p2" start="#t2" end="#t5" ana="#lid3"/>
  <fs lipDescr xml:id="lid3">
   <f name="part"><symbol value="corners"/></f> <f name="source"><symbol value="up"/></f>
   <f name="goal"><symbol value="down"/></f></fs>
  <kinesic type="handMove" xml:id="hand1" who="#p2" start="#t2" end="#t7" ana="#had1"/>
   <fs handDescr xml:id="hand1">
    <f name="part"><symbol value="shoulder"/></f>
    <f name="involvement"><symbol value="both"/></f>
    <f name="direction"><symbol value="up"/></f>
    <f name="shape"><symbol value="line"></f>
    <f name="velocity"><symbol value="medium"/></f></fs>
```

model when he understands the corresponding aspect of the meaning of a dialogue utterance. For instance, head nods may have different pragmatic meanings, such as expressing agreement or understanding, signalling turn acceptance or turn grabbing, or giving a positive answer (see Petukhova and Bunt, 2010b). Additionally, nonverbal communicative behaviour that emphasizes or supports the intended meaning of synchronous verbal behaviour is captured in terms of *qualifiers* that can be associated with a communicative function (e.g. uncertain, angry, happy, or anxious), resulting in more accurate descriptions of a speaker's behaviour (see Petukhova and Bunt, 2010a).

Dialogue act annotation is the assignment of functional meaning to stretches of dialogue behaviour. The unit in dialogue that carries a functional meaning is the *functional segment*, defined as a minimal stretch[15] of behaviour that has a communicative function (Geertzen et al., 2007). This definition implies that the identification of functional segment boundaries cannot be an independent process: segmentation and annotation on this view are simultaneous, rather than consecutive processes. Note also that functional segments may be discontinuous, may overlap, may stretch over more than one turn, and may contain material con-

[15]The rule is: do not include material in a functional segment which does not contribute to its communicative function(s).

tributed by different speakers.

The ISO 24617-2 taxonomy of communicative functions distinguishes 9 dimensions, addressing information about a certain task (the *Task* dimension); the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*). Identifying meaningful dialogue segments by considering multiple dimensions simultaneously results in very accurate description of the intended meaning of dialogue utterances (see illustrative example in Figure 1, and Petukhova and Bunt, 2011). A multidimensional approach to segmentation and annotation moreover supports the identification of relevant dialogue segments not only per dimension but also per modality, and the identification of complex multimodal multifunctional segments. We will see below how these can be represented according to the ISO 24617-2 standard.

6. DiAML representation

ISO 24617-2 includes the specification of the XML-based Dialogue Act Markup Language DiAML for the representation of dialogue act annotations. This representation relies on a three-level architecture:

1. the level of primary data, which may for example be a speech recording, a written text, or a video clip;
2. the marking of functional segments either directly in the primary data, in a coding of it, or in a lower-level representation of the primary data, such as the output of a tokenizer or action classifier for body movements;
3. the annotation associated with a functional segment.

At level 1, the primary data can be encoded in accordance with the TEI guidelines (TEI P5, 2007). For example, for the dialogue fragment of Figure 3, the speech turns and movements can be transcribed with timing information and a specification of the speaker as in (2).

Annotation in terms of type of body movements can be represented using the `@subType` attribute, as in (3):

```
(3) <kinesic type="gazeShift"
      xml:id="gaze1" subType="direct"/>
    <kinesic type="shoulderGesture"
      xml:id="shg1" subType="shrug"/>
    ...
    <kinesic type="lipMove" xml:id="lips3"
      subType="corner-down"/>
```

At level 2, functional segments can be identified by `functionalSegment` elements, which group together the components of multimodal communicative behaviour that constitute a multimodal functional segment. The example in Figure 3 of participant p2 turning his gaze to participant p1 (*gaze1*) and then averting it (*gaze2*), while producing the speech segment *Maybe* (`vec2`), performing a shoulder-shrug (*hag1*), constricting the forehead muscles (*fh1*), raising eyebrows (*brow1*), widening the eyes (*eye1*), lowering the lip corners (*lip3*) and raising the chin (*chin1*), can be represented as in (4):

```
(4) <fs type="functionalSegment"
     xml:id="fs1">
     <f name="verbalComponent" fVal="#vec2"/>
     <f name="gazeComponent" fVal="#gaze2"/>
     <f name="gestComponent" fVal="#hag1"/>
     <f name="headComponent" fVal="#head1"/>
     <f name="forehComponent" fVal="#fh1"/>
     <f name="eyebrComponent" fVal="#brow1"/>
     <f name="eyesComponent" fVal="#eye1"/>
     <f name="lipsComponent" fVal="#lip3"/>
     <f name="chinComponent" fVal="#chin1"/>
    </fs>
```

At level 3, in the DiAML representation of the dialogue act annotations the `@target` attribute, which can denote any pointer reference, is used to point to the multimodal functional segment. Example (5) illustrates the use of DiAML for the dialogue fragment in Fig. 3, containing two multimodal functional segments, corresponding to two dialogue acts:

```
(5) <diaml xmlns:=
      "http://www.iso.org/diaml/">
    <dialogueAct xml:id="da1" target="#fs1"
      sender="#p1" addressee="#p2"
      communicativeFunction"="suggestion"
      dimension="task"/>
    <dialogueAct xml:id="da2" target="#fs2"
      sender="#p2" addressee="#p1"
      communicativeFunction=
        "addressSuggestion"
      dimension="task"
      functionalDependence="#da1"/>
    </diaml>
```

Note that the DiAML annotation contains only semantic information; the description of the functional segments is not part of the annotation, but of the coding.

7. Conclusions and Outlook

In this paper we have described an approach to multimodal dialogue act annotation, starting from the conceptual distinction between the 'coding' of observable multimodal dialogue behaviour and the 'annotation' of such behaviour in semantic and pragmatic terms, and supported by experimental results in human and automatic multimodal dialogue coding and annotation. We provided XML representations both of multimodal coding and of multidimensional annotation of dialogue behaviour, showing how dialogue act annotations can be attached to multimodal data and how dialogue act annotations can be related to XML representations of multimodal functional segments.

In the near future we intend to extend this study in two directions. First, we will apply action recognition software that has recently been developed at Vicomtech, which is based on a robust approach to action unit tracking, segmentation and classification. The output is a sequence of time ordered action units that will be compared with manually performed codings in order to improve the automatic feature selection and classification. With the help of this new

software, we plan to produce a corpus of automatically transcribed and annotated AMI data. Second, with the corpus data obtained in this way we plan to perform experiments in automatic multimodal dialogue act recognition, from which we expect to gain a deeper understanding of the role of non-verbal communicative behaviour in dialogue.

8. References

Ahmad, M. and Lee, S.-W. (2006) Human Action Recognition Using Multi-View Image Sequences Features. *Proc. Intern. Conf. on Automatic Face and Gesture Recognition*, pp. 523-528.

Allen, J. and Core, M. (1997) Draft of DAMSL: Dialog Act Markup in Several Layers. See http://www.cs.rochester.edu/research/cisd/resources/damsl/

Allwood, J., Cerrato, L., Dybkjær, L., Jokinen, K., Navarretta, C., and Paggio, P. 2004. The MUMIN multimodal coding scheme. See http://sskkii.gu.se/jens/publications/bfiles/B80-3.pdf

AMI Consortium. (2005) Coding Guidelines for Individual Actions Annotation of the AMI Corpus, v 1.5.

Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In: S.-A. Jun (ed.) *Prosodic Typology - The Phonology of Intonation and Phrasing.* Oxford U. Press.

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

Bunt, H. (2006) Dimensions in Dialogue Act Annotation. In: *Proceedings of LREC 2006.*

Bunt, H. (2009) The DIT^{++} taxonomy for functional dialogue markup. *Proc. AAMAS Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, Budapest, 13-24.

Bunt, H., Kipp, M., and Petukhova, V. (2012). Towards an integrated scheme for semantic annotation of multimodal dialogue data. In *Proceedings of LREC 2012*, Istanbul.

De la Torre, F., Simon, T., Ambadar, Z. and Cohn, J.F. (2011) FAST-FACS: A computer-assisted system to increase speed and reliability of manual FACS coding. In: *Affective Computing and Intelligent Interaction (ACII).*

De la Torre, F. and Cohn, J.F. (2011) *Facial Expression Analysis, Guide to Visual Analysis of Humans: Looking at People*, Springer, Berlin.

Dornaika, F. and Davoine, F. (2006) On Appearance Based Face and Facial Action Tracking. In: *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9), 1107-1124.

Dornaika, F. and Raducanu, B. (2009) Simultaneous 3D Face Pose and Person-Specific Shape Estimation from a Single Image Using a Holistic Approach. In: *Proc. IEEE Int. Workshop on Applications of Computer Vision.*

Geertzen, J., V. Petukhova and H. Bunt (2007) A multidimensional approach to utterance segmentation and dialogue act classification. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 140-149.

Gut, U. and Looks, K. and Thies, A. and Gibbon, D. (2003) CoGesT: Conversational Gesture Transcription System, Version 1.0. Technical Report. Bielefeld University.

ISO (2010) ISO 24612:2010 Language resource management: Linguistic annotation framework (LAF), ISO, Geneva.

ISO (2010) Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO DIS 24617-2, Geneva, ISO, Geneva.

Jovanovič , N. (2007) *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings.* PhD Thesis, U. Twente.

Kang, H., Lee, C. W. and Jung, K. (2004) Recognition-Based Gesture Spotting in Video Games. In: *Pattern Recognition Letters* 25(15), pp. 1701-1714.

Kipp, M. (2004) *Gesture Generation by Imitation - From Human Behaviour to Computer Character Animation*, Boca Raton, Florida: Dissertation.com

Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H. (2006) Towards a common framework for multimodal generation: The behaviour markup language. In: *Proceedings of IVA-02.*

Kranstedt, A., Kopp, S., and Wachsmuth, I. (2002) MURML: A multimodal utterance representation markup language for conversational agents. In: *Proceedings AAMAS Workshop Embodied Conversational Agents, Autonomous Agents and Multi-Agent Systems* (AAMAS02), Bologna.

Krippendorff, K. (1995) On the reliability of unitizing continuous data. *Sociological Methodology* 25, 47 - 76.

Lösch, M. (2006) Selection of activities for key experiments. Institute of Computer Science and Engineering, University of Karlsruhe, Internal Report.

Lösch, M.,Schmidt-Rohr, S. R., Knoop, S. and Dillmann, R. (2008) Feature Selection for Human Activity Recognition Using Feature Taxonomies and User Comments. In: *Proceedings of the International Conference on Cognitive Systems*.

Liu, G.; Zhang, J., Wang, W. and McMillan, L. (2006) Human Motion Estimation from a Reduced Marker Set. In: *Proceedings of the Symposium on Interactive 3D Graphics and Games.*

Petukhova, V. and Bunt, H. (2010a) Introducing Communicative Function Qualifiers. In: *Proceedings Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, pp. 123-131.

Petukhova, V. and Bunt, H. (2010b). Towards an integrated scheme for semantic annotation of multimodal dialogue data. In *Proceedings of LREC 2010*, Malta, pp. 2556- 2563.

Petukhova, V. and Bunt, H. (2011). Incremental dialogue act understanding. In *Proceedings of IWCS 2011*, Oxford, pp. 235-244.

Ramanan, D. and Forsyth, D. A. (2003) Automatic Annotation of Everyday Movements. In: *Proceedings of the Neural Information Processing Systems Conference*.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. (2011) Real-time Human Pose Recognition in Parts from Single Depth Images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Stiefmeier, T. and Roggen, D. (2007) Gestures Are Strings: Efficient Online Gesture Spotting and Classification Using String Matching. In: *Proceedings of the International Conference on Body Area Networks*.

TEI (2007) Guidelines of Electronic Text Encoding and Interchange, edition P5. Text Encoding Initiative, Charlottesville, Virginia.

Unzueta, L. and Goenetxea, J. (2010) Cyclic and Non-cyclic Gesture Spotting and Classification in Real-Time Applications. In: *Proceedings of the Conference on Articulated Motion and Deformable Objects,* LNCS 6169, Springer, Berlin, pp. 172-181.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N., Kipp, M., Kopp, S., Mancini, M., Masella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., van Welberge, H., and van der Werf, R. (2007) The behaviour markup language: Recent development and challenges. In: *Proceedings of the IVA-07 Conference of Virtual Autonomous Agents.*

Zhou, F., De la Torre, F. and Hodgins, J. (2011) Hierarchical Aligned Cluster Analysis for Temporal Clustering of Human Motion. Under review for *IEEE Transactions on Pattern Analysis and Machine Intelligence.*