Towards an integrated scheme for semantic annotation of multimodal dialogue data

Volha Petukhova and Harry Bunt

Tilburg Center for Creative Computing
Tilburg University, Netherlands
{v.petukhova,harry.bunt}@uvt.nl

Abstract

This paper investigates the applicability of existing dialogue act annotation schemes, designed for the analysis of spoken dialogue, to the semantic annotation of multimodal data, and the way a dialogue act annotation scheme can be extended to cover dialogue phenomena from multiple modalities.

1. Introduction

In natural communication, the participants use all modalities that are available to them. Nonverbal behaviour is an essential part of human communication. This includes the use of gestures, facial expressions, gazes, posture shifts, etc; communicative resources which make the communication richer in many ways.

Several corpora with multimodal data transcriptions have in recent years become available, such as the AMI meeting corpus[1], the IFA Dialog Video corpus[2] and the ISL meeting corpus (Burger et al., 2002). They are largely used to study aspects of verbal and nonverbal behaviour in natural human conversations in general, e.g. interactive styles, emotionally relevant behaviour, affected speech, etc., as well as numerous aspects of human natural dialogue such as turn-taking behaviour, grounding, social regulating mechanisms and so on. Recent years witness a growing interest in the use of multimodal data for modelling tasks such as automatic interpretation and generation of multimodal communicative behaviour that dialogue participants naturally exhibit in interactive discourse.

For analysing the visual modality several coding schemes have been designed, in most cases based on the analysis of nonverbal actions in terms of behavioural low-level features. For example, the Facial Action Coding System (FACS)[3] codes facial expressions describing muscular activities that produce changes in facial appearance. HamNoSys[4] is a transcription system to code hand gestures by describing shapes, direction, speed, length and form of movement, hands orientation and location.

Few annotation schemes attempt to code semantic and pragmatic information in visual expressions. For example, the SmartKom Coding scheme is based on intentional information provided by a gesture (Steininger, 2001), such as command (interactional gesture), searching (supporting gesture), and emotions (residual gesture). DIME-DAMSL (Pineta et al., 2005) extends the DAMSL dialogue act annotation scheme (Allen and Core, 1997) with the annotation of the graphical modality that involves, for instance, pointing to, moving or adding a piece of furniture, or showing a catalogue. The MUMIN annotation scheme (Allwood et al., 2004) was developed for the study of gestures and facial displays in interpersonal communication and puts emphasis on the communicative function of such expressions, in particular their feedback and turn-managing functions.

Dybkjær and Bernsen in (2002), giving a comprehensive overview of coding schemes for multimodal data, point out that the majority of these schemes are designed for a particular purpose and are used solely by their creators. Standardisation has been achieved to some extent for coding behavioural features for certain nonverbal expressions, e.g. for facial expression, however, for the semantic annotation of such expressions combined with other modalities such as speech there is still a long way to go.

Existing dialogue act annotation schemes[5], however, are limited to analysis of spoken modality.

Over the last few years there has been increasing collaborative effort across research groups working on Embodied Conversational Agents, ECAs, to define a common framework for designing ECA systems, called SAIBA[6]. The AAMAS workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts' in 2008 and 2009 gathered researchers for first broad discussions about the issues surrounding the definition of a standard Functional Markup Language (FML)[7]. A major concern is that of dialogue acts. The relevance of the taxonomies that have been proposed in the literature and the way these can be used, adapted and extended for the ECA domain is discussed. It was con-

[1] Augmented Multi-party Interaction (http://www.amiproject.org/).

[2] http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/

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

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

[5] We analyzed 18 well-known dialogue act annotation schemes: DAMSL, SWBD-DAMSL, LIRICS, DIT^{++}, MRDA, Coconut, Verbmobil, HCRC MapTask, Linlin, TRAINS, AMI, SLSA, Alparon, C-Star, Primula, Matis, Chiba and SPAAC.

[6] Situation, Agent, Intention, Behavior, Animation framework specifies multimodal generation at a macro-scale. For more information please visit http://www.mindmakers.org/projects/SAIBA

[7] Detailed information can be found at http://hmi.ewi.utwente.nl/conference/EDAML

cluded that existing dialogue act taxonomies, such as Conversation acts (Allen et al., 1994), the Verbmobil coding scheme (Alexandersson et al., 1998), DAMSL (Allen and Core, 1997) and DIT (Bunt, 1999) can be used to further the development of FML. It was emphasised that the key difference between the coverage of most of these schemes and ECAs that ECAs communicate through a combination of verbal and nonverbal means, which means that for most of these schemes certain extensions are required.

In the context of the ISO project 24617-2 "Semantic annotation framework, Part 2: Dialogue acts", which aims to design a standard for annotating dialogues with dialogue act information, an approach has been developed for dealing with phenomena relating to multiple modalities used in dialogue, which are explored and motivated in this paper.

In this paper we outline two exploratory studies investigating the dialogue act annotation of multimodal data (Section 2). Sections 4 and 5 report the results and discuss their impact on dialogue act annotation scheme design. Finally, conclusions are drawn in Section 6.

2. Exploratory annotation study

2.1. Dialogue act annotation scheme

We used the DIT^{++}[8] dialogue act annotation scheme for the semantic annotation of multimodal dialogue data. This choice was motivated by several considerations. First of all, we wanted a taxonomy which has a well-defined near standard inventory of communicative functions with fine-grained distinctions, based on solid theoretical and empirical grounds. DIT^{++} is a starting point for the ISO project 24617-2 "Semantic Annotation Framework, Part 2: Dialogue acts", which aims at developing a standard for the markup of communicative functions in dialogue. DIT^{++} incorporates theoretical and empirical findings from other approaches (see Petukhova and Bunt, 2009c and Bunt and Schiffrin, 2007 for comparative analyses).

Second, we wanted a dialogue act taxonomy that allows to describe not only task-oriented communicative actions, but also actions related to other communicative dimensions such as feedback, turn taking, time management and dealing with social obligations, since many nonverbal acts are performed for purposes other than the underlying task. The DIT^{++} taxonomy distinguishes 10 dimensions, addressing information about: the domain or task (*Task*), processing status of the speaker (*Auto-feedback*) or partners (*Allo-feedback*), managing difficulties in the speaker's contributions (*Own-Communication Management*) or those of partners (*Partner Communication Management*), the speaker's need for time to continue the dialogue (*Time Management*), establishing and maintaining contact (*Contact Management*), allocation of turns (*Turn Management*), the way the speaker is planning to structure the dialogue (*Dialogue Structuring*), and social conventions (*Social Obligations Management*).

Third, a dialogue act annotation scheme should contain open classes, allowing suitable additions of those communicative functions which are specific for a certain modality. The DIT^{++} tagset contains 3 open classes.

Finally, a dialogue act annotation scheme should offer segmentation strategies that are flexible enough to identify meaningful dialogue units from multiple modalities. In natural conversation the use of speech is combined with nonverbal signs and vocal sounds, and all participants are most of the time performing some nonverbal communicative activity. DIT^{++} allows multiple segmentation. Communicative functions can be assigned in multiple dimensions to units called functional segments, which are defined as the functionally relevant minimal stretches of communicative behaviour (see Geertzen et al., 2007). Figure 1 illustrates the segmentation and annotation of multimodal units.

3. Corpus material and annotations

We conducted two annotation studies where annotators were asked to annotate dialogues with the DIT^{++} tagset using the ANVIL tool[9]: (1) using only speech transcription and sound; (2) using speech transcription, sound and video provided with transcriptions of nonverbal signals (gaze, head, facial expression, posture orientation and hand movements).

In both studies we used two scenario-based dialogues with a total duration of 51 minutes from the AMI corpus[10] . The transcriptions contain manually produced orthographic transcriptions, including word-level timings, and transcriptions of visible movements for each participant, including gaze direction; head movements; hand and arm gestures; eyebrow, eyes and lips movements; and posture shifts. Transcribers were asked to code low-level features such as form of movement (e.g. head: nod, shake, jerk); hands: pointing, shoulder-shrug, etc.[11]); direction (up, down, left, right, backward, forward); trajectory (e.g. line, circle, arch); size (e.g. large, small, medium, extra large); speed (slow, medium, fast); and repetitions (up to 20 times). The floor transfer offset (time difference between the start of a turn and the end of the previous turn) and duration of a movement (in milliseconds) were computed automatically.

We examined agreement between annotators in labelling communicative functions using Cohen's kappa measure (Cohen, 1960). Two experienced annotators reached substantial agreement (kappa = .76).

We compared the annotations with respect to the number and nature of (1) functional segments identified; (2) communicative functions altered; (3) communicative functions specified; and (4) communicative functions assigned to single functional segments.

4. Study results

The analysis showed that nonverbal communicative behaviour may serve four purposes:

[8] For more information about the tagset, please visit: http://dit.uvt.nl/

[9] For more information about the tool visit: http://www.dfki.de/~kipp/anvil

[10] See http://www.amiproject.org/

[11] Hand gesture transcription was performed according to Gut,U., Looks, K., Thies, A., and Gibbon, D. (2003). CoGesT: Conversational Gesture Transcription System. Version 1.0. Technical report. Bielefeld University http://www.spectrum.uni-bielefeld.de/modelex/publication/techdoc/cogest/

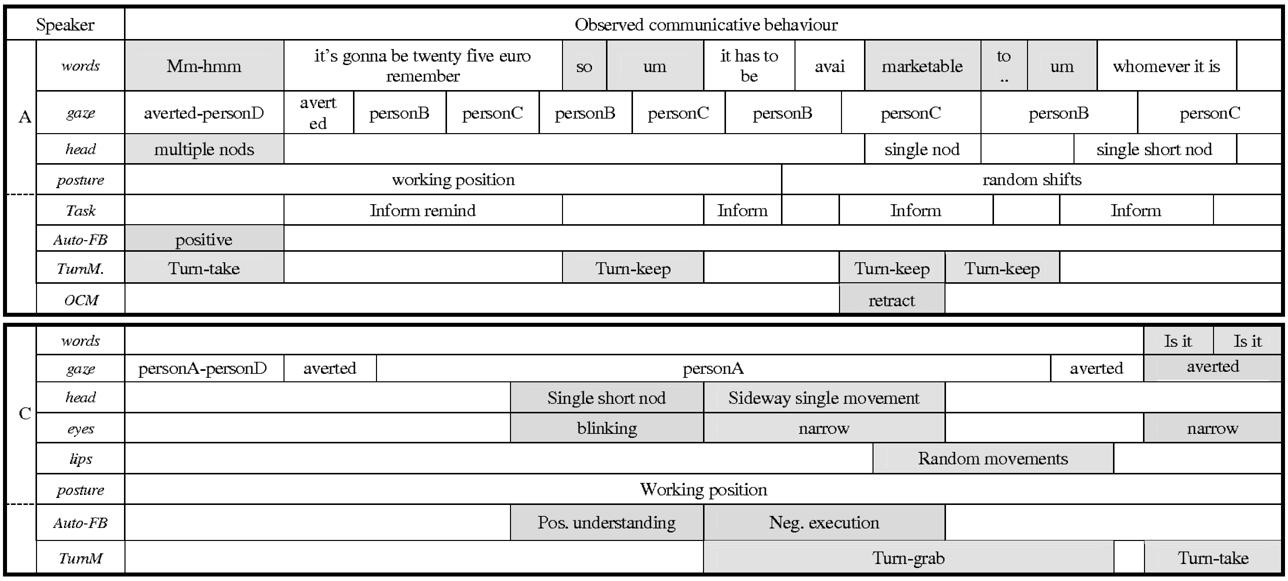
Speaker
Observed communicative behaviour
A
words
Mm-hmm
it's gonna be twenty five euro remember
so
um
it has to be
avai
marketable
to ..
um
whomever it is
gaze
averted-personD
avert ed
personB
personC
personB
personC
personB
personC
personB
personC
head
multiple nods
single nod
single short nod
posture
working position
random shifts
Task
Inform remind
Inform
Inform
Inform
Auto-FB
positive
TurnM.
Turn-take
Turn-keep
Turn-keep
Turn-keep
OCM
retract
C
words
Is it
Is it
gaze
personA-personD
averted
personA
averted
averted
head
Single short nod
Sideway single movement
eyes
blinking
narrow
narrow
lips
Random movements
posture
Working position
Auto-FB
Pos. understanding
Neg. execution
TurnM
Turn-grab
Turn-take

Figure 1: Transcription, multidimensional segmentation and annotation.

1. emphasizing or articulating the semantic content of dialogue acts;
2. emphasizing or supporting the communicative functions of synchronous verbal behaviour;
3. performing separate dialogue acts in parallel to what is contributed by the partner;
4. expressing a separate communicative function in parallel to what the same speaker is expressing verbally.

4.1. Full-fledged dialogue acts

Our study shows that the number of identified functional segments is larger when taking the visual modality into account in addition to the speech. In the first study 1,917 functional segments were identified when annotating speech only. In the second study annotators identified 2,396 functional segments when using both speech and nonverbal signals, i.e. about 20% more.

The 479 new functional segments, which have only nonverbal components, form a single full-blown dialogue act or multiple dialogue acts. These acts mainly address auto-feedback (68.5%): positive (65.3%), negative (3.2%). Signs of feedback notably overlap the main speaker's utterance (850ms on average). They are used frequently around the utterance boundaries: (1) in final boundary position in 39.4% of the cases; (2) near the start of a new segment after speaker identification or continuation signals like discourse markers (e.g. 'so', 'and', 'because', 'such as', 'but'); editing expressions; restarts; or retractions, in 22.3% of all cases; (3) during turn-internal hesitation phases (36% of all cases).

4.1.1. Feedback

In face-to-face communication, where the communicative partners have permanent visual feedback, they also use the modalities other than speech to indicate the flow of their understanding process. First of all, if the speaker addresses some particular person in the conversation, they usually have direct eye contact. Gaze direction clearly shows others where the focus of attention lies. Also, listeners usually shift their posture slightly (e.g. leaning forward, backward or aside, shifting one's weight in a chair) to communicate a potentially changing emotional state and an awareness of surrounding activity or tension. The interlocutor can indicate his attention by side-way turns of his head to the left or right. Positive feedback was realized by several types of head nods and jerks. The face can also display the state of cognitive processing, such as perception and interpretation, e.g. half lowered eyebrows, half opened mouth. Often head nods are combined with a smile or laugh and eye blinking. We also observed that intensification of the up-down head movements (repeated short nods) accompanied by some posture shifts (leaning forward) indicates that the interlocutor is positive about the previous message and wants to add something, and therefore also wants to *take the turn*.

Negative feedback can be provided on various levels: attention, perception, interpretation, evaluation and execution. Negative attention is generally characterized by absence of any noticeable verbal or nonverbal activity of the dialogue participant or when the participant's focus of attention is directed to a dialogue partner other than the current speaker. The speaker, in such cases, may attract attention from his interlocutors by making pauses and looking at them, leaning to the intended addressee or making sharp hand movements. Negative feedback at the level of perception is often signalled by puzzled facial expression (curving the mouth downward, lowering the eyebrows and eyelids, dropping the jaw, constricting the forehead muscles), cupping the ear hand gesture (meaning 'I can't hear you'). Negative feedback at higher levels is signalled by head shakes (signalling opposition or inability to perform a requested action), and raising the shoulders (meaning 'I don't know' or 'Maybe'), waggles (head movements back and forth or left to right signalling uncertainty), lip-pout or compression (signalling disappointment, disbelief, disliking or disagreement), lowering eyebrows (indicator of skepticism, disagreement or doubt).

4.1.2. Turn Management

Of all dialogue acts performed nonverbally 4.7% are used for the purpose of managing the allocation of turns. In other words, dialogue participants perform certain actions to take the turn over or to give the turn away. In Turn Management, a distinction is made between turn-obtaining acts (turn-initial acts) and acts for keeping the turn or giving it away (turn-final acts). A turn-initial function indicates whether the speaker of this turn obtains the speaker role by grabbing it (*turn grab*), by taking it when it is available, (*turn take*) or by accepting the addressee's assignment of the speaker role to him (*turn accept*). A turn ends either because the current speaker assigns the speaker role to the addressee (*turn assign*), or because he offers the speaker role without putting any pressure on any particular addressee to take the turn (*turn release*). A turn may also have smaller units with boundaries where a reallocation of the speaker role might have occurred, but does not occur because the speaker indicates that he wants to keep the turn. Such a segment has a *turn keep* function.

Dialogue participants do not just start speaking if they want to say something or stop speaking if they want to end their contribution. As (Petukhova and Bunt, 2009b) showed, participants in dialogue signal that they want to have the turn often by using gaze re-directions, gaze aversion, facial expression and posture shifts. In multi-party conversations gaze plays a more significant role in managing fluent turn transitions than in two-person dialogues, because of the increased uncertainty about who will be the next speaker. As for gaze patterns that accompany turn-initial segments, in 29.4% of the cases the participant has direct eye contact with his addressee. In 11.8% of the cases the participant who wants to have the next turn gazes at more than one of the partners, most probably verifying their intention concerning the next turn. A dialogue participant who aims for the next turn first gazes at one or more partners, and averts his gaze shortly before starting to speak (44.1%).

Comparable patterns were observed in previous studies. A speaker usually breaks mutual gaze while speaking and returns gaze to the addressee upon turn completion (Kendon, 1967). Goodwin in (1981) claims that the speaker looks away at the beginning of turns and looks towards the listeners at the end of the turn. More recently, Novick (1996) found that 42% of the turn exchanges follows a pattern in which the speaker looks toward the listener while completing the turn. After a short moment of mutual gaze the listener averts his gaze and begins the next turn.

Independent from the possible meanings of specific types of head movements, and from their feedback functions, head movements are used for turn management purposes. It was noticed in (Hadar et al., 1984) that speakers use head movements to mark syntactic boundaries and to regulate the turn-taking process. In our data the intention to have the next turn was signalled by repetitive short head nods (positive feedback), by waggles (indicated negative feedback or uncertainty), or head shakes (signalling disagreement).

Hand and arm gestures that may be related to the participant's intention to have the turn were not observed frequently. We identified some shoulder shrugs that signalled uncertainty accompanied by head waggles and hand movements when a participant listening to the speaking partner suddenly moves his hand/fist away from the mouth or makes an abrupt hand gesture for acquiring attention.

To signal the intention to have the next turn, participants frequently made random silent lip movements, compressing, biting, licking, or pouting their lips. They also often keep their mouth (half-) open, narrow (possible sign of negative feedback) or widen (indicating surprise) their eyes accompanied by lowering or raising eyebrows.

Various types of upper-body posture shifts were often used as turn-initial signals. Participants would change their body orientation from working position (both hands on the table, leaning slightly forward, head turned to the speaker) to leaning forward, backward or aside, producing random shifts (shifting one's weight in a chair), shifting from bowing position (bending, curling, or curving the upper body, usually while writing). Cassell et al. in (2001) looked at posture shifts at turn boundaries and discourse segment boundaries, and showed that both boundaries had an influence on posture shifts. Posture shifts with the upper body were found more frequently at the start of a turn than in the middle or end (48%, 36%, and 18% respectively).

Several studies showed that participants is dialogue use semantic, syntactic, pragmatic, prosodic, but also visual features to signal turn endings (Ford & Thompson, 1996; Grosjean & Hirt,1996; De Ruiter et al., 2006; Barkhuysen et al., 2008, among others). We observed that the speaker giving the turn away (either releasing or assigning) has direct eye contact with the communicative partner(-s). Turn assigining events are often signaled by deictic head nods or deictic hand gestures. Releasing the turn the speaker usually terminates any hand gesticulation. Orientation of the upper-body is mainly towards the potential next speaker. Turn keeping, by contrast, is signalled by a significant amount of gaze aversion, by holding/freezing the currently produced hand gesture or by a palm-down stopping gesture (meaing 'wait' or 'hold on') preventing others interrupt the speaker. Speakers used to stay in the working position (both hands on the table, leaning slightly forward, head turned to the addressee), lean forward or to the addressee, randomly shift their posture (shifting one's weight in a chair) and bow (bending, curling, or curving the upper body, usually while writing or searching notes).

4.1.3. Discourse Structuring

About 2% of all nonverbally performed dialogue acts are used for the purpose of dialogue structuring. Topic shifts were announced by raising a hand or a finger and palm-down gesture. Emphasizing head nods used for this purpose mean also that everything up to now was processed successfully and the speaker is ready to move to the next discussion topic. Establishing mutual gaze and positioning the upper body in the working positions or breaking mutual gaze and leaning backward, respectively, were used for opening and closing the dialogue.

Modality	Verbal expressions	Vocal /prosody	Gaze direction	Head movement	Facial expression	Gesture	Posture orientation

Uncertainty	may (not)
	might (not)
	could (not)
	should (not)
	probably(not)
	(un)likely
	maybe(not)
	'not sure'
	'you know?'
	'I guess', etc.
		high standard deviation in pitch;
		voice breaks;
		jitter;
		shimmer;
		filled/ unfilled pauses;
			aversion redirection
			involuntary eye movements
				waggles	lip-compression;
					lip-pout;
					biting/liking;
					lowering eyebrows;
					constricting forehead muscles
						adaptors, e.g. self-touching;
						shoulder shrug
							posture shift

Certainty	shall
	will(not)
	can(not)
	would(not)
	must(not)
	certainly(not)
	definitely(not)
		low standard deviation in pitch;
		no pauses
		no restarts
			direct eye contact;	head nod (for emphasis)	thin lips;
					pushing up the chin boss;
					widely open eyes;
						beat gestures	leaning forward /to addressee

Table 1: Expressions of modality

4.1.4. Time Mangement

24.8% of all nonverbal acts were assigned the communicative function of *stalling* (time management). Gaze aversion within an utterance was interpreted by annotators as indicating stalling. Head waggles are observed as stalling signals, as are various types of self-touching: touching, scratching, or holding the back of the neck or head with the open palm and rubbing the cheek or side of the neck.

Our investigation shows that the nonverbal 'speaker-' and 'listener-regulatory' movements described above have functions for managing speaking turns and time, providing feedback and structuring the conversation, or a combination of those. These movements can be interpreted in a particular context with or without the co-occurring spoken utterance, and can be analysed in terms of dialogue acts. Indeed, our comparative analysis of 18 existing dialogue annotation schemes (see Introduction for the list) shows that these communicative aspects are reflected in a significant majourity of them: Feedback is not defined only in Linlin and Primula; Turn management acts are not defined in HCRC MapTask, Verbmobil, Linlin, Alparon and C-Star; Discourse Structuring is not defined in TRAINS and Alparon; and Time Management is not defined in MRDA, HCRC MapTask, Linlin, Maltus, Primula and Chiba.

4.2. Communicative function alteration and specification

In a number of cases the communicative functions assigned to speech segments were corrected after annotators got access to visual signals. Mostly, this concerned an adjustment of the level of feedback, e.g. from understanding to evaluation or execution (6.8%). In (Petukhova and Bunt, 2009a) we noticed that participants in dialogue provide different types of evidence to their partners if they merely understand the partner's intentions than if they also adopt the information provided (positive execution feedback). Several types of head movement were studied, and the features were investigated that were used to interpret these signals as indicating understanding or agreement. It was shown that dialogue participants use multiple signals and modalities to provide evidence of grounding at different levels, and that conversational partners perceive and understand this more accurately when they can rely on multiple information sources. While simple head nods were perceived as a signal of successful understanding, more complex expressions, such as a combination of multiple slow head nods with lip movements and blinking, were perceived as signals of belief transfer (adoption). Also words like 'uh-uh' in combination with head nods were interpreted by multiple judges as an understanding signal, whereas variations of 'yes' accompanying head nods were seen as signals of adoption.

Dialogue participants often express assessments of the validity of their propositions. Kendon (2004) observed that nonverbal acts which are not part of the propositional or referential meaning of the utterance may have modal functions, e.g. indicating whether the speaker regards what he is saying as a hypothesis or as an assertion. About 47% of all functional segments in our data are modalized (34.5% uncertain, 12.6% certain). A degree of certainty can be expressed verbally, e.g. by auxiliary verbs like 'may', 'might', 'could'; adverbs like 'probably', 'likely'; and by expressions like 'I guess', 'I'm not sure', as well as nonverbally, e.g. waggles, shoulder shrugs, puzzled look. Table 1 gives an overview of observed expressions.

It was further noticed that nonverbal expressions may reveal the speaker's attitude towards the addressee(-s), towards the content of what he is saying, or towards the actions he is considering to perform. For instance, nonverbal expressions are used to mark new, important information, or mark out logical components of that. As such they reinforce verbal communication and allow to accentuate or emphasise words or ideas. To stress the importance of information that the speaker is providing he can use beat gestures, which are known to accompany important information, as well as eyebrow movements to indicate where the focus of the addressee's attention should be positioned. Along with hand and eyebrow movements speakers often use head nods for emphasis coinciding with the most prominent words in an utterance. For example:

(1) wording: *I'm gonna do an opening*
head: ....*nod*

Moreover, nonverbal acts may signal a speaker's emotional or cognitive state. Pavelin (2002) calls these nonverbal expressions *modalizers* or *modal* gestures. We observed the following attitudes and emotions in our data: assertive, thinking or reflecting, surprised, confused, amused, sceptical, interested, disappointed, and guilty. Attitudes and emotions are often communicated by face.

We conclude here that nonverbal expressions are often used to support the communicative functions of the synchronous verbal behaviour and help to disambiguate it (e.g. understanding vs adoption). Nonverbal signs also emphasize or qualify a communicative function expressed by a verbal utterance, expressing modality (certain vs uncertain) or attitudinal and emotional state. No existing dialogue act annotation scheme deals with this type of information. A proposal to this effect is formulated in Section 5.

4.3. Multifunctionality in multimodal utterances

A verbal functional segment has on average 1.3 communicative functions (also confirmed in Bunt, 2009), whereas we observed that using information from all modalities gives 1.4 functions per segment on average. Table 2 presents the relative frequency of co-occurrences of multiple functions in various dimensions.

In spite of the fact that the average number of functions per segment does not differ much, multimodality is significant for enabling the multifunctionality of utterances in some dimensions. For instance, our observations show that nonverbal communicative acts are very often concerned with feedback and other interaction management dimensions. Speech-focused movements, for example, accompanying relatively unpredictable content words (e.g. iconic gestures during lexical search), body-focused movements (e.g. searching for an elusive word or expression in memory) normally indicate that the speaker needs some time to gather his/her thoughts or to formulate the utterance, and is therefore stalling for time, but also keeping the turn. Sometimes pauses increase the pressure on other participants to say something. The longer the pause, the more pressure builds on the other person(-s) to respond. Pauses near the beginning of an utterance can have the function of contact check, requesting attention. Speakers often make short pauses until the gaze of a recipient has been obtained and secured.

Repetitive head nods, lip movements, raising a finger, and beginning gesticulation may indicate that the previous contribution has been understood and the participant would like to grab the turn to add or correct something. Nonverbal expressions which are used to manage turns may also be used by the speaker to edit and structure his own speech. It was noticed by Butterworth (Butterworth, 1980), for example, that an excessive amount of gaze aversion when the speaker is having difficulty formulating a message may lead a listener to interfere. Here, also expressions of uncertainty (e.g. lip compression, curving the mouth downwards, lowering eyebrows and eyelids, constricting the forehead muscles, head waggles) may invite the partner to take the turn and assist.

Nonverbal signals also add functionality to utterances addressing the Task dimension, in particular for Turn Management, Time Management, Discourse Structuring, Own Communication Management and Allo-Feedback. The speaker who is ready with his dialogue contribution and wants the addressee to take the turn may signal this by gazing at the potential next speaker or by pointing at the addressee or performing a deictic head movement. When the speaker shifts his posture, e.g. leaning or turning to an addressee, near the end of his turn, he signals that he expects the addressee to react to what the speaker just said. If the speaker gazed at more than one participant near the end of an utterance this often means that the speaker wants to release the turn and somebody else to continue the dialogue. Gaze aversion, by contrast, often means that the speaker has not yet finished his contribution and wants to keep the turn. In this case, if the speaker experiences difficulties in formulating his utterance or needs some time to gather his thoughts, he may signal this by clems (involuntary eye movements), self-touching gestures like rubbing checks or neck, touching hair or biting lips, or by using some iconic gesture modelling an object in the air.

It was observed that feedback utterances like ’okay’ or ’yes’ in combination with repetitive head nods often have a communicative function of turn take or grab, and also abrupt head movements signal an effort by the listener to take the speaking turn. Head nods as signs of positive feedback are used for structuring the discourse when the speaker indicates that what was happened in dialogue up to now is successfully processed and his is ready to move to the next topic, or close the discussion or entire conversation.

Interesting nonverbal behaviour was observed with respect to speaker speech production and editing (own communication management). The speaker interrupts himself by speech corrections or editing and indicates that he wants to delete part of an utterance and/or substitute this by something else. Retractions frequently occur at the beginning of an utterance and within other hesitations and phrasal breaks. When the speaker’s gaze reached a non-gazing participant or the partner’s gaze arrived later with some delay, the speaker often restarted or retracted his utterance indicating by this that he wishes to gain the addressee’s attention. Such behaviour is multifunctional in the sense that the speaker signals that he corrects or retracts his utterance and by doing this he indents to elicit feedback from his interlocutors at the same time.

Dialogue act taxonomies that take the multifunctionality of utterances into account such as DIT^{++}, LIRICS, DAMSL, MRDA and Coconut, known as *multidimensional* dialogue act annotation schemes, allow assigning multiple functional labels to one stretch of communicative behaviour and therefore provide a better account for the multifunctionality of verbal and nonverbal signals.

4.4. Articulating semantic content

Our studies show that about 39% of all transcribed nonverbal signals neither contribute to the communicative function of a verbal utterance nor form a full-fledged dialogue act on their own. Nonverbal signals are often used for articulating the semantic content of a dialogue act, relating to the propositional or referential meaning of an utterance.

Hand gestures are often used deictically when the speaker points to entities which he is referring to. This accounts for about 48.6% of all hand movements. For example:

(2) wording: *Press this little presentation*
hand: ....*point*................

within	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task		1.1 (1.2)	0.1 (2.7)	5.6 (8.5)	2.6(3.4)	0.3(0.3)	0(0)	4.3(4.6)	0.3(0.3)	1.5(1.5)
Auto-F.	0.5(0.7)		0(0)	12.7(15.5)	0.5(2.6)	0.3(3.1)	0(0)	0(0)	0(0)	0(0.5)
Allo-F.	0(3.3)	0(0)		23.7(23.7)	1.2(1.5)	0(0)	0(0)	0(15.4)	0(5.1)	0(0)
Turn M.	39.3(40.8)	6.2(12.2)	1.8(6.0)		49.6(60.6)	0.7(1.1)	0(0.3)	2.5(5.9)	0(0.7)	0.4(0.7)
Time M.	34.6(41.7)	0.5(3.5)	0(11.2)	9.1(9.7)		0(0.5)	0(0)	0(4.2)	0(1.4)	0(0.6)
DS	1.7(6.8)	0(6.8)	0(0)	6.7(20.9)	0(1.7)		0(0)	0(1.7)	0(0)	0(8.4)
Contact M.	0(0)	0(0)	0(0)	18.2(18.2)	0(0)	0(0)		0(0)	0(0)	0(0)
OCM	77.9(80.9)	0(0)	0(5.4)	6.5(6.5)	0(8.0)	0(0.9)	0(0)		0(0)	0(0)
PCM	0(0)	0(0)	0(18.2)	27.3 (27.3)	0(0)	0(0)	0(0)	0(0)		0(0)
SOM	0.9(0.9)	0(1.2)	0(0)	1.2(8.3)	0(1.2)	13.9(13.9)	0(0)	0(0)	0(0)	

Table 2: Co-occurrences of communicative functions across dimensions in % excluding nonverbal expressions and including nonverbal expressions (in brackets).

Some iconic, metaphoric and pantomimic gestures were observed which form part of the semantic content or specify the semantic content of an utterance (51.4%). For example:

(3) *wording: then we'll move into acquaintance including a tool training exercise*
hand: ....semi-sphere..

(4) *wording: then we've moved to age group twenty five to thirty five*
hand: ....away-motion................

(5) *wording: The younger group of people would want smaller*
hand: ....size(both hands)

The speaker in example (3) performs a metaphoric gesture 'a spherical shape' accompanying the word 'including'. The speaker in (4) performs a pantomimic gesture depicting the movement. The speaker in (5) performs an iconic gesture using both hands with open palms towards each other depicting size of the object in question.

These types of nonverbal acts normally co-occur with speech, i.e. are coverbal. Their meaning is inseparable from the meaning of verbal components and their interpretation arises directly from the overall speech utterance. Therefore, such nonverbal acts, which can be considered as pure semantic acts, as a rule do not have a communicative function on their own, but together with the speech determine the meaning of the multimodal utterance as a whole.

5. Extensions to DIT^{++}

Our studies show that the DIT^{++} inventory of communicative functions was unable to cover all phenomena exhibited by participants' nonverbal behaviour. People may be less straightforward in expressing their communicative intentions. They often indicate their attitude toward their communicative partners and toward what they are saying. They emphasize, express doubts, criticize, show interest and so on. In this respect some fine-tuning is required.

Adding communicative functions like 'Amused Suggestion' and 'Uncertain Answer' would lead to an explosive growth of the tagset and would probably not provide a complete solution anyway. Instead, we propose to add a set of qualifiers that can be attached to communicative function in order to describe the speaker's behaviour more accurately. For instance, *modal* qualifiers can be introduced to annotate the strength of the speaker's beliefs about the validity of a proposition, having the values *uncertain* and *certain*. For more fine gradation of uncertainty one might introduce additional values like certain negative ("definitely not") - uncertain negative ("probably not") - uncertain positive ("probably") - certain positive ("definitely"). Emotional and attitudinal phenomena in dialogue can be labeled with different levels of granularity: coarse (positive, negative and neutral); medium (basic emotions comparable to Ekman's 6 emotions), and fine (labels for specific emotions like misery, annoyed, worry; specific attitudes like critical, impatient, agreeable, serious, curious). We propose to leave this category open-ended, allowing specific qualifiers to be chosen according to the needs of particular applications and tasks.

qualifier attribute	qualifier values	CF category
modality	uncertain, certain	info-providing functions
mode	angry, happy, surprised, ...	info-providing functions; feedback functions
conditionality	conditional, unconditional	action-discussion functions
partiality	partial, complete	responsive functions; feedback functions

Table 3: Function qualifier attributes, values, and function categories

Other qualifications of communicative functions could be *conditionality*, referring to the possibility (with respect to ability and power), necessity or volition of performing an action, and can therefore only be attached to action-discussion functions; and *partiality* that limits the scope of a communicative function in addressing only part of the semantic content of the utterance to which the current utterance is related.

Table 3 summarizes the qualifier attributes and values that we propose, indicating in the rightmost column the categories of communicative functions to which they may be attached.

6. Conclusions

In this paper we explored the role of nonverbal modalities in the interpretation of dialogue behaviour and investigated experimentally whether it is possible to apply a dialogue act annotation scheme for the semantic annotation of multimodal data. The general conclusion of our experiments is that a well-worked out, fine-grained, open multidimensional dialogue act taxonomy such as DIT^{++} (but also DAMSL, MRDA or Coconut) is usable for this purpose when some adjustments are made in order to deal with the modal, attitudinal and emotional information that is transmitted by nonverbal modalities. We proposed a solution for adding these aspects to a dialogue act annotation scheme without changing its set of communicative functions, in the form of qualifiers that can be attached to communicative function tags.

7. Acknowledgments

This research was conducted within the project Multidimensional Dialogue Modelling, sponsored by the Netherlands Organisation for Scientific Research (NWO), under grant reference 017.003.090. We also very thankful to anonumous reviewers for their valuable comments.

8. References

Alexandersson, J., et al. 1998. *Dialogue acts in Verbmobil-2*. Second edition. Report 226. DFKI Saarbruecken, University of Stuttgart; TU Berlin; University of Saarland.

Allen, J. et al. 1994. *The TRAINS Project: a case study in building a conversational planning agent.* TRAINS Technical Note 94-3. University of Rochester.

Allen, J., Core, M. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers*.

Allwood, J., Cerrato, L., Dybkjær, L., Jokinen, K., Navarretta, C., and Paggio, P. 2004. *The MUMIN multimodal coding scheme*. Technical report.

Barkhuysen, P., Krahmer, E., and Swerts, M.. 2008. The interplay between auditory and visual cues for end-of-utterance detection. The Journal of the Acoustical Society of America, 123(1):354–365.

Bunt, H. 1999. *Dynamic interpretation and dialogue theory*. In: M. Taylor, F. Neel, and D. Bouwhuis (eds.), *The structure of multimodal dialogue II*, Amsterdam: Benjamins, pp. 139–166.

Bunt, H. and Schifrin, A. 2006. *Methodological aspects of semantic annotation*. In: Proceedings LREC 2006, Genova.

Bunt, H., and Schiffrin, A. 2007. *Documented compilation of semantic data categories*. LIRICS Deliverable D4.3 available at `http://lirics.loria.fr/`

Bunt, H. 2009. *Multifunctionality and multidimensional dialogue semantics*. Proceedings of the DiaHolmia Workshop on the Semantics and Pragmatics of Dialogue, Stockholm, Sweden, pp. 3–15.

Burger, S., MacLaren, V., and Yu, H. 2002. *The ISL Meeting Corpus: the impact of meeting type on speech style*. Proceedings ICSLP, Denver CO, USA.

Butterworth, B. 1980. *Evidence from pauses in speech*. In B.Butterworth(ed.), Language Production: Speech and Talk, vol.1. Academic Press, London, pp. 155–177.

Cassell, J., Nakano, Y.I., Bickmore, T.W., Sidner, C.L., and Rich, R. 2001. Non-Verbal Cues for Discourse Structure. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France.

Clark, H. and Krych, M.A. 2004. *Speaking while monitoring addressees for understanding*. Journal of Memory and Language, 50: 62-81.

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

Dybkjær, L. and Bernsen, N.O. 2002. *Data Resources and Annotation Schemes for Natural Interactivity: Purposes and Needs*. Proceedings of the LREC2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation.

Ford, C.E., and Thompson, S.A. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: Emanuel A. Schegloff and Sandra A. Thompson, editors, Interaction and grammar, Cambridge: Cambridge University Press, pp. 135– 184.

Geertzen, J., Petukhova, V., and Bunt, H. 2007. *A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification*. Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, pages 140–149.

Goodwin, C. 1981. Conversational Organization: Interaction between hearers and speakers. New York: Academic Press.

Grosjean, F. and Hirt, C. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. Language and Cognitive Processes, 11:107–134.

Hadar, U., Steiner, T.J., Grant, E.C., and Rose, F.C. 1984. The timing of shifts of head postures during conversations. *Human Movement Science*, 3:237–245.

Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologia*, 26: 22–63.

Kendon, A. 2004. *Gesture: visible action as utterance*. Cambridge University Press, Cambridge.

Novick, D.G., Hansen,B., and Ward, K. 1996. Coordinating Turn-taking with Gaze. In: *Proceedings of the International Symposium on Spoken Dialogue*, Philadelphia, PA, pp.53–56.

Pavelin, B. 2002. *Le Geste à la parole*. Toulouse: Presses Universitaires du Mirail.

Petukhova, V., and Bunt, H. 2009. *Grounding by nodding*. Proceedings of the 1st Conference on Gesture and Speech in Interaction, Poznan, Poland.

Petukhova, V., and Bunt, H. 2009. *Who's next? Speaker-selection mechanisms in multiparty dialogue*. Proceedings of the DiaHolmia Workshop on the Semantics and Pragmatics of Dialogue, Stockholm, Sweden, pp. 19–26.

Petukhova, V., and H. Bunt. 2009. *Dimensions in communication*. TiCC Technical Report TR 2009-002, Tilburg University.

Pineda, L., Castellanos, H., Coria, S., Estrada, V., López, F., López, I., Meza, I., Moreno, I., Pérez, P.,and Rodríguez, C. 2005. *Balancing Transactions in Practical Dialogues*. Technical report, Department of Computer Science, Mexico.

de Ruiter, J.-P. Mitterer, H., and Enfield, N.J. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. Language, 82: 515–535.

Steininger, S. 2001. Labeling Gestures in SmartKom - Concept of the Coding System. Technical Report, LMU Munich.