Grounding by nodding

Volha Petukhova and Harry Bunt

Tilburg Center for Creative Computing
Tilburg University, the Netherlands

v.petukhova@uvt.nl; harry.bunt@uvt.nl

Abstract

This paper addresses the question what aspects of a dialogue participant's behavior are perceived as evidence of grounding and at which level of information exchange: that of understanding or that of agreement. Our observations show that a range of verbal and nonverbal expressions are used to signal correct understanding or adoption of the partner's beliefs. Head movements are known to be signals showing a participant's state of cognitive processing, e.g. agreement, disbelief, or lack of understanding. Nods, in particular, which vary in speed, duration, timing, and intensity may convey different meanings. We found that, analyzed in isolation, head nods do not enable an adequate interpretation of the participant's state of grounding; they have to be considered in combination with other signs in order to allow successful interpretation as grounding acts of a particular type.

1 Introduction

To be successful, participants in dialogue have to coordinate their activities on many levels. In the speaker role, a participant not only produces utterances but also evaluates whether the addressee(-s) attend to, perceive, understand, and react to the speaker's intentions. An addressee's task is to attempt to understand the speaker's utterances, react to their intentions, and report on his processing. The coordination of the beliefs and assumptions of the participants is a central issue in any communication, the basic coordination problem being that of building shared or mutual beliefs out of individual ones. A set of propositions that the dialogue participants mutually believe is called their *common ground*, and the process of establishing and updating the common ground is called *grounding*. While 'common ground' is not directly observable, grounding mechanisms are accessible through observable dialogue behavior, e.g. evidence of understanding what is said in dialogue is provided by feedback acts. The nature of such evidence depends on the communicative situation. In face-to-face conversation, for example, participants may present evidence of grounding through body movements and gaze re-direction, while in telephone conversations only verbal and vocal signals are available for the participants.

Nonverbal means play an important role in the grounding process in face-to-face dialogue. For example, eye gaze is the most basic form of showing attention to what the speaker is saying, and head nods have a communicative function of acknowledgement signaling that the previous utterance was understood, without necessarily signaling acceptance (Clark 1996). Goodwin (1981) notices that dialogue participants utilize both their bodies and a variety of vocal phenomena to show each other the type of attention and, reciprocally, the type of orientation they expect from others. For example, the speaker makes pauses and restarts his utterance when his gaze reaches a non-gazing recipient, or when late-arriving gaze of a recipient reaches a gazing speaker, or when

recipient movements are noticeably delayed. Novick et al. (1996) found that the proportion of mutual gaze during conversational difficulties is greater at turn boundaries than within the turn. Nakano et al. (2003) observed that maintaining gaze on the speaker is interpreted as evidence of non-understanding, requesting additional information (73% of all cases); by contrast, continued gaze on task-related objects (e.g. looking on a map) is interpreted as evidence of understanding (52% of all cases).

All these findings suggest that nonverbal communicative means contribute especially to lower levels of grounding, signaling attention, perception and understanding of each other's communicative actions. As grounding may occur at many (if not at all) levels of processing, one would expect evidence of grounding to also be provided at many levels, including higher ones such as evaluation and the adoption of beliefs. We show that this certainly happens in the case of complex nonverbal signs such as combinations of head nods, gaze re-direction and facial expressions. Such nonverbal evidence of higher-level grounding is observed in empirical data and also successfully recognized by multiple judges.

This paper is organized as follows. In Section 2 we briefly present various views on grounding, including the approach using the framework of Dynamic Interpretation Theory (DIT) proposed by Bunt et al. (2007), which we use for our investigations. In Section 3 we present results from our analyses of dialogue data and perceptual experiments. Section 4 draws conclusions.

2 Grounding

Several models of grounding have been proposed in the literature. One of the best known is the Contribution Model (Clark and Schaefer, 1989). According to this model participants in dialogue perform collective actions ('contributions') that result in grounding. To make a contribution requires (1) *content specification* (a speaker tries to specify the content of his contribution, and the partners try to register that content), and (2) *grounding* (participants attempt to establish the mutual belief that they understand what was said). Each contribution has two phases: a *presentation* phase, where the speaker presents an utterance for the addressee to consider, and an *acceptance* phase, where the addressee gives evidence that he believes he understands what the speaker means by this utterance. Evidence of understanding includes verbatim repetitions of part of the previous utterance, acknowledgements (*'uh uhu', 'yeah'*), initiation of a relevant next contribution, or letting the speaker proceed with his utterance, indicating satisfaction with the partner's presentation.

Traum (1999) points out some weaknesses of the Contribution Model. It is difficult to determine whether a particular utterance is part of a presentation or an acceptance phase, and how to measure enough acceptance to consider the previous contribution(-s) as grounded. The Contribution Model does not specify what mutual beliefs are created and when, and how they are updated. The computational model of grounding proposed by Traum (1994) makes use of *grounding acts* that have a specific function in advancing the mutual understanding. To model the multi-utterance exchanges necessary for mutual understanding, Traum proposed *discourse units* which consists of an initial presentation and as many utterances as needed to make this act mutually understood. Traum's computational model does not go beyond the grounding of an utterance and as such only models mutual belief about understanding. Grounding is considered as the process of establishing the mutual understanding of each other's intentions and actions, not that of the utterance content. The Contribution Model requires that participants specify the content, but does not provide means to represent the content of contributions. Neither model computes the semantic content of an utterance to specify what information is being or has to be grounded.

Bunt et al. (2007) propose their view on grounding from a semantic perspective using the framework of Dynamic Interpretation Theory (DIT) (Bunt 2000). *Dialogue context* and *dialogue acts* are the main ingredients of this model. The dialogue context is partly dynamic, in the sense of changing during a dialogue as the result of the participants interpreting each other's communicative behavior, reasoning with the outcomes of these processes, and planning further activities. Dialogue acts are defined as operators that update contexts in certain ways, which can be described by the *communicative function* and the *semantic content* of that dialogue act. The semantic content

(propositional, referential) corresponds to what the utterance is about. Communicative functions are defined as specifications of the way semantic content is to be used by an addressee to update his information state when he understands the utterance. This gives a formal semantics to the notions of communicative function and semantic content. Information is transferred from one dialogue participant to another through belief creation (*understanding*) and belief transfer (*adoption*). An utterance is understood by the addressee when the addressee comes to believe that the preconditions of an intended dialogue act hold. For example, if A requests B to perform an action then the understanding of A's request will be that B believes that A wants B to perform an action, and that A assumes that B is able to perform this action. Not only the correct understanding is needed for the grounding of this request, but also evidence of believing. If B reacts as *'Yes, of course'*, then A may be expected to believe that B plans to perform the requested action. This is called the *adoption* of information.

To be sure that information is indeed transferred, a speaker needs evidence of correct understanding of his communicative behavior and of being believed. In face-to-face interaction speakers receive such evidence through verbal and nonverbal expressions. The example in Figure 1 shows that different nonverbal and verbal expressions and their combination may convey different meanings. In this example, B says *"but I th I think regardless we're we're aiming for the under sixty five"*. To come believe that *p* (*'we are aiming for the under sixty five'*), B should get evidence that A, C and D understand his utterance and believe its content *p*. The first head movement of speaker A in combination with gaze directed to B signals his understanding of speaker B's intention to have the turn; A's and D's multiple short head nods signal their understanding of B's intention to continue as a speaker (*'I think…'*). A's utterances *'Under sixty five'*, *'Okay'* and *'That's a good start'* accompanied by multiple short nods provide evidence of understanding (and positive evaluation) but not of adoption, since A offers that proposition for further debate. Thus, B believes that A believes that B believes that *p*, but B does not yet know whether A believes that *p*. The evidence of understanding and adoption is provided by speaker C when he uses gaze directed to B, long double nods (where the first one most probably indicates understanding (and is also a turn taking act since B by his gaze invites C to participate in the dialogue) accompanied with single eye blinking and verbal *'Yep'* to express agreement with B's inform. Thus, B believes that C believes that B believes that *p* and B weakly believes that C believes that *p* is true. In the grounding model of Bunt et al. (2007), these beliefs may be strengthened by continuing dialogue when both have evidence that both know that both believe that *p*.

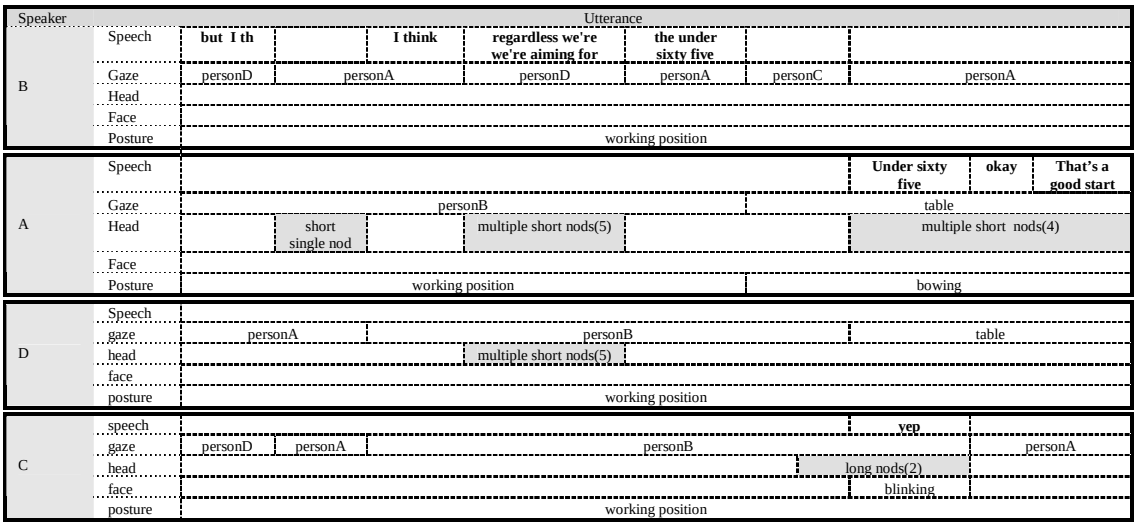

Figure 1: Example of multimodal utterances from the AMI corpus

Therefore, as we see in the example presented in Figure 1, some evidence given nonverbally is about understanding and its interpretation does not lead to belief transfer, whereas other nonverbal signals may be interpreted as successful belief adoption. In the next section we examine which

types of nonverbal expressions and their combinations can be interpreted as adoption signals and which merely signal understanding. This will be investigated by means of perception experiments with multiple judges.

3 Experiment

3.1 Stimuli and procedure

We used human-human multi-party interactions in English from the *AMI corpus*[1], which contains manually produced orthographic transcriptions for each individual speaker, including word-level timings. The meetings are video-recorded and provided with sound files.

The nonverbal behavior of the dialogue participants was transcribed using video recordings for each individual participant, running the videos without sound to eliminate the influence of what was said. The transcription includes gaze direction, head movements, hand and arm gestures, eyebrows, eyes and lips movements, and posture shifts. Transcribers were asked to annotate low-level behavioral features such as *form of movement* (e.g. head: nod, shake, jerk, etc.; hands: pointing, shoulder-shrug, etc.[2]; eyes: blinking, widen, etc.; lips: compress, flatten, (half)open, etc.), *direction* (e.g. up, down, left, right, etc.), *trajectory* (e.g. line, circle, etc.), *size* (e.g. large, small, medium, etc.), *speed* (slow, medium, fast) and *repetitions* (up to 20 times). The *floor transfer offset* (fto: difference between the time that a turn starts and the moment the previous turn ends) and *duration* (in milliseconds) were computed. At this stage no meaning was assigned to movements.

Speech and nonverbal signs were annotated with the DIT++ tagset[3] using the ANVIL tool[4]. From the annotated data we randomly selected 60 video clips with 6 different speakers (3 male, 3 female). All six meeting participants were English native speakers (3 speakers of American English and 3 of British English).

The duration of each clip was about 10 seconds and contained the full turns of the previous speaker and the current speaker. 16 naïve subjects (4 male and 12 female, all between the ages of 20 and 40) participated in the perception experiments. They were given the task to answer the question whether they think that a participant understands the dialogue act of the previous speaker or that he/she agrees with the previous speaker. Subjects had 10 seconds to react to each stimulus and were allowed to watch every video as many times as they liked.

3.2 Results

First, inter-subject agreement was examined using Cohen's kappa measure (Cohen 1960)[5]. The judges reached a substantial overall agreement rating the stimuli (overall kappa 0.68). They recognized the dialogue participant behavior as signals of belief adoption better than those of correct understanding, reaching a higher agreement (kappa scores of 0.9 and 0.54 respectively).

Next we determined nonverbal features that might be helpful for explaining why a participant's behavior was interpreted as an expression either of correct understanding or of belief adoption. The following features were investigated:

- wording of an utterance, if any (and for the most frequent words like *'yeah'* and *'uh-uhu'*);
- gaze (to person, table, slides, or averted);
- head movement, if any (nods or jerks) and for these:
    - number of repetitions;
    - duration;
    - floor transfer offset;
    - speed (number of movements per second);
    - size (extra small, small, medium, large, extra large);

[1] **A**ugmented **M**ulti-party **I**nteraction (http://www.amiproject.org/).
[2] Hand gesture transcription was performed according to Ulrike Gut, Karin Looks, Alexandra Thies and Dafydd Gibbon (2003). CoGesT: Conversational Gesture Transcription System. Version 1.0. Technical report. Bielefeld University.
[3] For more information about the tagset, please visit: http://dit.uvt.nl/.
[4] ANVIL is free for research purposes. For download information visit http://www.dfki.de/~kipp/anvil
[5] This measure of agreement takes chance agreement into account and has the following interpretation: 0=None; 0-0.2=Small; 0.2-0.4=Fair; 0.4-0.6=Moderate; 0.6-0.8=Substantial; and 0.8-1.0=Almost Perfect.

- eyebrow movement, if any;
- eye shape change (e.g. blinking, widen, narrow), if any;
- lips movement, if any;
- hand movement, if any;
- posture shift, if any;
- some combinations of these features.

We performed Pearson's correlation tests and measured for each class label the correlations between the proportion of judges that chose this label and the numerical features described above. Table 1 presents the correlation results for the 'adoption' label (the correlation coefficient values for the 'correct understanding' label are the opposite ones).

It is observed that if the dialogue participant combined head nods with verbal elements, especially the use of *'yeah'*, this was perceived by evaluators as a signal of belief adoption, see e.g. the behavior of speaker C in the example in Figure 1. Combination of *'uh-uhu'* and head nods is more ambiguous; no significant correlation was observed.

Signs of understanding are usually produced more silently. The speaker usually signals that he has understood the contribution without showing his acceptance or agreement with the partner. Understanding utterances notably overlap the main speaker's utterance (average fto = -850ms). 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).

Expressions of belief adoption, by contrast, are used around turn boundaries and may slightly overlap the main speaker utterance (average fto = -54ms).

Head nods were mostly interpreted as adoption/agreement signals, and jerks (single backward head movement) as signals of understanding. The number of head nods positively correlates with the agreement interpretation: the more nods, the more probable that the speaker is adopting the partner's beliefs. Moreover, slow multiple head nods were interpreted by most of the judges as signals that partner beliefs are adopted.

Feature	Pearson's R
head nod(-s) + wording	.55* (p=0.000)
head nod(-s) + 'yeah'	.43* (p=0.000)
head nod(-s) + 'uh-uhu'	.2 (p=0.123)
duration	.17 (p=0.186)
floor time offset	.34* (p=0.07)
speed of movements	.22 (p=0.07)
size of movements	.027 (p=0.834)
number of repetitions	.25* (p=0.045)
head nod	.29* (p=0.02)
head jerk	-.29* (p=0.02)
gaze pattern 'person-averted'	.47* (p=0.06)
blinking	.25* (p=0.49)
eyebrows movement	.012 (p=0.925)
lips movements	.42* (p=0.001)
hand movements	.039 (p=0.762)
posture shift	-.16 (p=0.210)
fast single nod	-.13 (p=0.305)
fast multiple nods	.13 (p=0.32)
slow single nod	-.025 (p=0.847)
slow multiple nods	.37* (p=0.003)

Table 1: Features correlated with the proportion of votes for 'adoption' (* differs significantly from zero according to two-sided t-test, t < .05)

As for gaze pattern, when agreeing with their partners speakers exhibit certain regularities in the gaze behavior that accompanies their head nods. They first look at the partner and avert their gaze near the end of the agreement phrase.

Distinctive for agreement utterances were head nods in combination with lips movements, the speaker either flattening the lips (the mouth appears to be longer than usual in the horizontal plane, with lips compressed against the teeth) or smiling (lips corner-up and elongated). The test results also show that dialogue participants when expressing agreement with their partners often perform head nods together with eye blinking. Thus, head movements, which are diverse in form, speed, number of repetition, timing and accompanying verbal and nonverbal signs, convey different meanings and therefore play a different role in grounding processes.

4 Discussion

In this study we used the DIT model of grounding in dialogue, which views information exchange as occurring through understanding and believing each other. We assumed that dialogue participant would provide different types of evidence to their partners if they merely understand the partner's intentions then if they also adopts the information provided. We studied several types of head movements that correlate with understanding and adoption, and investigated the features of understanding or adopting behaviors which are used to interpret these singals. We showed that dialogue participants use multiple signals and modalities to provide grounding evidence at different levels, and that conversational partners perceive and understand each other's intention more accurately when they can rely on multiple information sources.

A point for future research is to investigate whether the costs of grounding in face-to-face conversation are lower than in dialogues where participants do not have direct eye contact, and whether this depends on the task. Clark and Brennan (1991) notice that partners in a collaborative task monitor and coordinate their behavior to minimize their collective effort as well as the costs that arise in joint activity, therefore they should benefit from the possibility to have visual contact. The so-called *gaze advantage hypothesis* suggests a benefit in performance when partners can share gaze. For visual search tasks (e.g. direction-giving dialogues like MapTask) it was found that people are twice as fast and efficient when they use shared gaze than when they don't (Nakano et al. 2003; Brennan et al. 2008). The situation may be different for other types of tasks such as negotiation tasks, problem-solving, or non-collaborative tasks. It would also be interesting to investigate whether grounding can be achieved entirely nonverbally in situations with severe limitations on the use of speech.

Bibliography

Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., and Zelinsky, G.J. 2008. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. Cognition, 106(3). pp. 1465-1477

Bunt, H. 2000. Dialogue pragmatics and context specification. In: Bunt, H., Black, W.(eds.) Abduction, Belief and context in Dialogue. Amsterdam: Benjamins. pp.81–150.

Bunt, H., S. Keizer, and Morante, R. 2007. A computational model of grounding in dialogue. In: Proceedings of the Workshop in Discourse and Dialogue. Lecture Notes in Computer Science 4629. Antwerp, Belgium. pp. 591-598.

Clark, H., Schaefer, E. 1989. Contributing to discourse. Cognitive Science 13. pp. 259–294

Clark, H.H., Brennam. 1991. Grounding in communication. In: Resnock, L.B., Levine, J.M., and Teasley, S.D. (eds.) Perspective on Socially Shared Cognition. Washington, DC, USA: APA Books, pp. 127-149.

Clark, H.H. 1996. Using Language. Cambridge, UK: University Press.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Education and Psychological Measurement, 20. pp.37-46.

Goodwin, C. 1981. Achieving mutual orientation at turn beginning. In: Conversational organization: Interaction between speakers and hearers. New York: Academic Press. pp. 55-89.

Nakano, Y.I., Reinstein, G., Stocky, T., and Cassell, J. 2003. Towards a model of face-to-face grounding. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 553-561.

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.

Traum, D. 1994. A Computational Theory of Grounding in Natural Language Conversation. PhD Thesis. Dep. of Computer Science, University of Rochester.

Traum, D.R. 1999. Computational models of grounding in collaborative systems. In: Brennan, S.E., Giboin A., and Traum D.R. (eds.) Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems. Menno Park, California. pp. 124-131.