Clarifying Intentions in Dialogue: A Corpus Study*

Julian J. Schlöder and Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
julian.schloeder@gmail.com, raquel.fernandez@uva.nl

Abstract

As part of our ongoing work on grounding in dialogue, we present a corpus-based investigation of intention-level clarification requests. We propose to refine existing theories of grounding by considering two distinct types of intention-related conversational problems: *intention recognition* and *intention adoption*. This distinction is backed-up by an annotation experiment conducted on a corpus assembled with a novel method for automatically retrieving potential requests for clarification.

1 Introduction

Dialogue is commonly modelled as a *joint activity* where the interlocutors are not merely making individual moves, but actively collaborate. A central coordination device is the *common ground* of the dialogue participants, the information they mutually take for granted (Stalnaker, 1978). This common ground is changed and expanded over the course of a conversation in a process called *grounding* (Clark, 1996). We are interested in the mechanisms used to establish agreement, *i.e.*, in the conversational means to establish a belief as *joint*. To investigate this issue, in this paper we examine cases where grounding (partially) fails, as indicated by the presence of clarifications requests (CRs). In contrast to previous work (*i.a.*, Gabsdil, 2003; Purver, 2004; Rodríguez and Schlangen, 2004), which has mostly focused on CRs triggered by acoustic and semantic understanding problems, we are particularly concerned with problems related to intention *recognition* (going beyond semantic interpretation) and intention *adoption* (*i.e.*, mutual agreement). The following examples, from the AMI Meeting Corpus (Carletta, 2007), are cases in point:

(1) A: I think that's all.
B: **Meeting's over?**

(2) A: Just uh do that quickly.
B: **How do you do it?**

(3) A: I'd say two.
B: **Why?**

In these examples, it cannot be said that B has fully grounded A's proposal, but also not that B rejects A's utterance. Rather, B asks a question that is conducive to the grounding process. In (1), B has apparently understood A's utterance, but is unsure as to whether A's intention was to conclude the session. We therefore consider CRs like B's question in (1) as related to *intention recognition*. In contrast, in (2) and (3), B displays unwillingness or inability (but no outright refusal) to ground A's proposal, and requests further information she needs to establish common ground, *i.e.*, to *adopt* A's intention *as joint*. Requests for instructions have also been related to clarification in Benotti's (2009) work on multiagent planning.

In this paper, we present a corpus-based investigation of intention-level clarification, part of an ongoing project that aims to analyse the grounding process beyond semantic interpretation. In the next section, we introduce some theoretical observations and refine existing theories of grounding (Clark, 1996; Allwood, 1995) by distinguishing between *intention recognition* and *intention adoption*. We then present a systematic heuristic to retrieve potential clarification requests from dialogue corpora and discuss the results of a small-scale annotation experiment.[1] We end with pointers for future work.

*The research presented in this paper has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 *ESSENCE: Evolution of Shared Semantics in Computational Environments* (http://www.essence-network.com/).

[1] We will make our annotated data freely available.

Proceedings of the 11th International Conference on Computational Semantics, pages 46–51,
London, UK, April 15-17 2015.

	Level		Joint Action	Example Clarification
1	contact		A and B pay attention to each other	*Are you talking to me?*
2	perception		A produces a signal and B perceives it	*What did you say?*
3	understanding		A conveys a meaning and B recognises it	*What did you mean?*
4.1	uptake	intention recognition	A intends a project and B understands it	*What do you want?*
4.2		intention adoption	A proposes a project and B accepts it	*Why should we do this?*

Table 1: Grounding hierarchy for speaker A and addressee B with refined uptake level.

2 Theoretical Observations

As extensively discussed by Hulstijn and Maudet (2006), the intentional level we are interested in is commonly denoted with the term *uptake*. In particular, in Clark's (1996) stratification of the grounding process into four distinct levels (see Table 1 for our take on it), the fourth level, "proposal and consideration (uptake)," is related to the speaker's intentions. When discussing joint projects at level 4, Clark introduces the notion of *joint construals*: the determination and consideration of speaker meaning, including the intended illocutionary force (Clark, 1996, pp. 212–213). However, he also points out that uptake may fail due to unwillingness or inability: "when respondents are unwilling or unable to comply with the project as proposed, they can *decline* to take it up" (Clark, 1996, p. 204). We contend that this difference between construal and compliance—between intention recognition and intention adoption—has been obscured in the literature so far.[2] For example, in their annotation scheme for CRs, Rodríguez and Schlangen (2004) reproduce the underspecification in labelling their level 4 CRs as "recognising or evaluating speaker intention."

Since we, with Clark (1996), consider such intentional categories to be part of the grounding hierarchy, we expect problems on an intentional level to be evinced in much the same way as other conversational mishaps: in particular by CRs aimed at fixing these different types of conversational trouble. When studying the CRs annotated as intention related in the corpus of Rodríguez and Schlangen (2004) we indeed find examples related to *recognition* and others which aim at *adoption*:[3]

(4) K: okay, again from the top
I: **from the very top?**
K: no, well, [. . .]

(5) K: for me that is in fact below this
I: **why below?**
K: yes, it belongs there, all okay.

In (4), speaker I has evidently not fully understood what K's question is, despite having successfully parsed and understood the propositional content of K's utterance. On the other hand, I displays no such problem in (5), but rather some reluctance to adopt K's assertion as common ground. We consider (4) to be a clarification question related to *intention recognition* whereas the one in (5) relates to *intention adoption*. A particularly striking class of intention recognition CRs are *speech act determination* questions as in the following example:[4]

(6) A: And we're going to discuss [. . .] who's gonna do what and just clarify
B: **Are you asking me whether I wanna be in there?**

Our hypothesis is that the classes of clarification requests related to intention recognition and intention adoption, respectively, are distinct and discernible. In particular, we propose to improve upon Clark's (1996) hierarchy by splitting his uptake-level into two, separating recognition from adoption. Table 1 shows our amended hierarchy and constructed examples for clarification requests evincing failure at a certain level. To test this hypothesis, we have surveyed existing corpora of CRs and assembled a novel corpus of intention-related CRs to check if annotators could reasonably discern the two classes.

[2]While DIT++ (Bunt, 2012) stratifies the grounding hierarchy into "attention / perception / interpretation / evaluation / execution," it is similarly underspecified: To us, evaluation (*e.g.*, checking an asserted proposition for consistency) relates to intention adoption, whereas (semantic) understanding and (pragmatic) intention retrieval (*e.g.*, recognising on level 4.1 that an indicative was intended as an *inform* act and hence requires a consistency check on level 4.2) are again distinct categories.

[3]We thank the authors for providing us with their annotated corpus; in the dialogues, I is explaining to K how to assemble a paper airplane. We had the German-language examples translated to English by a native speaker of German.

[4]Retrieved from the British National Corpus (BNC) (Burnard, 2000) using SCoRE (Purver, 2001).

3 Corpus Study

3.1 Previous Studies

Our work builds on previous corpus studies of CRs (Purver et al., 2003; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005). However, existent studies are not perfectly suited for investigating grounding at the level of intentions.[5] Firstly, the annotation scheme of Purver et al. (2003; 2004), which the authors apply to a section of the BNC (Burnard, 2000), makes use of semantic categories that cannot easily be mapped to the intention-level distinctions introduced in the previous section. Secondly, while the schemes employed by Rodríguez and Schlangen (2004) and Rieser and Moore (2005) (both based on Schlangen, 2004) do include a category for intention-level CRs, the corpora they annotate—the Bielefeld Corpus and the Carnegie Mellon Communicator Corpus, respectively—are highly task-oriented and hence the intentions of the interlocutors are to a large degree presupposed: the participants intend to fulfil the task. Finally, in all cases, the focus of the authors did not lie with intentional clarification and therefore they might have left out questions in their annotations that are interesting to us, in particular more complex intention adoption CRs (which may not have been considered CRs to begin with, given the lack of well established theoretical distinctions discussed in the previous section).

For our study, we have chosen to extract questions from the AMI Meeting Corpus (Carletta, 2007), a collection of dialogues amongst four participants role-playing a design team for a TV remote control. The dialogues are loosely task- and goal-oriented, but the conversation is mostly unconstrained. Due to this setting, we expect a larger amount of discussion and decision making, which should give rise to more intention-level CRs. In addition, the rich annotations distributed with the AMI Corpus enabled us to apply a sophisticated heuristic to automatically extract potential CRs, which we describe next.

3.2 Data

The AMI Corpus is annotated with dialogue acts, including a class of '`Elicit-*`' acts denoting different kinds of information requests/questions, but without specifically distinguishing CRs. However, the corpus is also annotated with relations between utterances, loosely called *adjacency pair* annotation,[6] which indicates whether or not an utterance is considered a direct reply to another one. We utilise observations on the sequential nature of CRs ("other-initiated repair") in group settings made by Schegloff (2000) to assemble a set of possible clarification requests as follows. Take all utterances Q where:

a. Q is turn-initial and annotated as an '`Elicit-`' type of dialogue act, spoken by a speaker B.

b. Q is the second part of an adjacency pair; the first part (the *source*) is spoken by another speaker A.

c. Q is the first part of another adjacency pair; the second part (the *answer*) is spoken by A as well.

This heuristic is based on the intuition that CRs are proper questions (*i.e.*, utterances that demand an answer) with a backward-looking function (*i.e.*, related to an earlier source utterance) that are typically answered by the speaker of the source. We expect this heuristic to have a sufficiently high recall to be quantitatively applicable, but are aware that it cannot find each and every CR.[7]

There are 338 utterances Q in the AMI Corpus satisfying the criteria above. We note that the annotation manual for the AMI Corpus states that CRs are usually annotated as '`Elicit-`' acts, but that some very simple CRs (*e.g.*, *'huh?'*) can instead be tagged as '`Comment-about-Understanding (und)`.' However, this class also contains some backchannel utterances: positive comments about understanding. If we apply the same heuristic to the utterances annotated as '`und`,' we find 195 additional possible CRs. We confirmed that our heuristic successfully separates CRs from backchannels, and that these CRs are

[5] We have carefully studied the annotated data described in Purver et al. (2003) and Rodríguez and Schlangen (2004), which was kindly provided to us by the authors upon request.

[6] See `http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf`.

[7] In particular, previous work indicates that some CRs are simply not answered; Rodríguez and Schlangen (2004) report 8.7% unanswered CRs in their corpus. Our heuristic does not find these.

indeed related to levels 1–3 of Clark's (1996) hierarchy. However, these utterances are not the primary subject of our study. We henceforth refer to CRs on levels 1–3 collectively as *low-level*.

3.3 Annotation Procedure

As indicated above, we are primarily interested in the 338 possible CRs annotated as '`Elicit-`' dialogue acts and therefore included only these in our annotation. Since our main interest is in intention-level CRs and our primary ambition is the investigation of intention adoption *vs.* intention recognition, we used the following simple annotation scheme: Each question found by our heuristic is annotated as one of {`not`,`low`,`int-rec`,`int-ad`,`ambig`}, where the categories are defined as follows.

- **not CR.** Select this category if you are sure that the question is not a clarification request. That is, if it does *not* serve to better the askers understanding of the previous highlighted utterance. For instance if the question is requesting novel information, moving the dialogue forward.
- **low CR.** Select this category if the question indicates that the asker has not fully understood the *semantic / propositional content* of the previous highlighted utterance. This includes, for example, *word meaning* problems, *acoustic* problems, or *reference resolution*.
- **intention recognition CR.** Select this category if the question indicates semantic understanding, but that the CR utterer *has not fully understood (or is trying to guess)* the speaker's goal/intention (the intended function of the previous highlighted utterance). The prototypical case is *speech act determination*.
- **intention adoption CR.** Select this category if the question indicates the CR utterer *has understood/recognised* the speaker's main goal (their intention), *but does not yet accept it because he wants/needs more information or he has incompatible beliefs*. For instance, if the CR utterer asks about the reason behind the speaker's utterance before accepting it, or requests information needed to carry out her proposal.
- **ambiguous.** Sometimes it may not be possible to decide what function a CR has precisely, maybe due to a lack of context. In those cases, annotate the question as ambiguous.

We instructed our annotators to follow a decision tree where they first decide whether a question is clearly *not* a CR, and only otherwise consider the different categories of CRs. This is because in a pilot study we found that the distinction between '`not CR`' and '`intention adoption CR`' was difficult for some annotators. To reduce the confusion, we defined the '`not CR`' class as only clear-cut cases of not-CR questions, at the risk of incurring a higher amount of ambiguity when the decision tree bottoms out, *i.e.*, when a question that was not definitely not a CR could not be matched to a CR-category after all. Our annotation scheme only refines one dimension (namely, '`source`') of the multi-dimensional schemes applied by Rodríguez and Schlangen (2004) and Rieser and Moore (2005). Since our main ambition in this work is to establish the two levels of intentionality, we leave a fuller annotation with further dimensions—such as syntactic categories like Schlangen's (2004) '`form`'—for future work.

Nevertheless, this is a difficult annotation task: Annotators can only play the role of overhearer and therefore have a more indirect access to the intentions of the interlocutors. In addition, CRs in particular can be fragmented and ambiguous. Therefore, annotators were shown a substantial dialogue excerpt starting 10 utterances before the source and ending with either the 10th utterance after the answer to the CR or with the CR-asker's next reply (the *follow-up*). We found that answer and follow-up are particularly helpful in determining the function of a CR: the answer gives hints towards the speaker's interpretation of the CR, and the follow-up can show whether the asker agrees with that construal.[8]

In the full study, the corpus was annotated by 2 expert annotators, since we deemed the task to be too complex and fine-grained for naïve annotators. One third of the corpus was annotated by both annotators, the remaining two thirds by one annotator each. To create a gold-standard on the overlapping segment, the annotators discussed the utterances where their initial judgement differed and mutually agreed on the appropriate annotation.

[8] Rodríguez and Schlangen (2004) include the CR asker's '`happiness`' (as evinced by the follow-up) in their annotation.

Category	Count	including 'und'	Example
not CR	90 (27%)	-	*A: 'You can call me Peter.' – B: 'And you are? In the project?'*
low-level	78 (23%)	273 (62%)	*A: 'Seventy-five percent of users find it ugly.' – B: 'The LCD?'*
intent. recognition	53 (16%)	53 (12%)	*A: 'I think that's all.' – B: 'Meeting's over?'*
intent. adoption	77 (23%)	77 (17%)	*A: 'That's a very unnatural motion.' – B: 'Do you think?'*
ambiguous	40 (12%)	40 (9%)	
Total	338 (100%)	443 (100%)	

Table 2: Distribution of clarification requests in our corpus with examples for each category.

3.4 Results

In the five-way classification task described above, our annotators had an agreement (Cohen's κ, 1960) of $\kappa = 0.76$ on the overlapping third of the corpus;[9] of $\kappa = 0.85$ in the boolean task of determining whether an utterance is a CR; and of $\kappa = 0.82$ in the boolean task of retrieving intention-related CRs from all other questions. The distribution of categories is shown in Table 2. In order to compare our distribution to previous work, we have also recorded the distribution we obtain when dropping the items annotated as '`not CR`' and adding the questions annotated as '`Comment-about-Understanding (und)`' as low-level CRs. Then the total number of CRs in our corpus is 443.

The AMI Corpus contains about 42,000 turns, so we found that roughly 1.1% of turns receive clarification according to our heuristic. Previous studies have indicated a higher number: Purver (2004) reports about 4% and Rodríguez and Schlangen (2004) about 5.8%. Rodríguez and Schlangen (2004) themselves conjecture that their corpus might contain an unusually high amount of CRs due to the setting (an instructor guiding a builder). For comparison, we have manually extracted CRs from a 2500-turn subset of the AMI Corpus: We found 52 CRs in that segment, indicating that about 2% of turns prompt a CR. It is to be expected that our heuristic misses some CRs, *e.g.*, ones that do not receive an answer, and its coverage is dependent on how systematic the adjacency pair annotation in the AMI Corpus is.

While our heuristic only retrieves an estimated 50% of CRs,[10] the distribution of classes we found is comparable to the results described by Rodríguez and Schlangen (2004) and Rieser and Moore (2005): They report 63.5% and 75%, respectively, of low-level CRs and 22.2% / 20% on intention-level. Rodríguez and Schlangen (2004) mark the remaining 14.3% as ambiguous, whereas Rieser and Moore (2005) report 5% "other/several" and do not mention an ambiguity class.[11] By and large, this is comparable to the distribution we found. We have low ambiguity (9%) compared to Rodríguez and Schlangen (2004) because we conflated different categories of lower-level CRs into one '`low CR`' category. As we had hoped, we find a larger amount (29%) of intention-level CRs than the previous studies. We take the similarity in distributions as tacitly confirming the viability of our heuristic for quantitative evaluation.

4 Conclusion

We have theoretically motivated a distinction within grounding hierarchies between *intention recognition* and *intention adoption* and have created a novel corpus of intention-level CRs to investigate its tenability. Our corpus is not only novel in its contents, but also in its construction: unlike previous studies, we have developed and applied a suitable heuristic that exploits rich existing annotations to automatically find possible clarification requests. A small-scale annotation experiment on our corpus showed that the theoretical distinction we propose is viable. Our immediate next step in this project is a deeper investigation into the form and problem sources of the intention-level CRs in our corpus, including a more fine-grained annotation.

[9] Rodríguez and Schlangen (2004) report $\kappa = 0.7$ in the task of determining the level of understanding that the CR addresses. However, their categorisation is different from ours. In particular, they do not include a '`not CR`' category.

[10] We surveyed the CRs not found by our heuristic and attribute this mostly to the adjacency pair annotation; however, in addition to CRs that are not answered at all, there are also CRs that are answered by a different person than the source speaker.

[11] Their category "ambiguity" refers to a class of CRs dubbed "ambiguity refinement" and not to uncertainty in the annotation.

References

Allwood, J. (1995). An activity based approach to pragmatics. *Gothenburg papers in theoretical linguistics* (76), 1–38.

Benotti, L. (2009). Clarification potential of instructions. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Bunt, H. (2012). The semantics of feedback. In *Proceedings of the 16th SemDial Workshop on the Semantics and Pragmatics of Dialogue (SeineDial)*.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation 41*(2), 181–190.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 1*(20), 37–46.

Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, Stanford, CA, pp. 28–35.

Hulstijn, J. and N. Maudet (2006, June). Uptake and joint action. *Cognitive Systems Research 7*(2-3), 175–191.

Purver, M. (2001, October). SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London.

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. Ph. D. thesis, King's College, University of London.

Purver, M., J. Ginzburg, and P. Healey (2003). On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pp. 235–255. Springer.

Rieser, V. and J. D. Moore (2005). Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rodríguez, K. J. and D. Schlangen (2004). Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th SemDial Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*.

Schegloff, E. A. (2000). When 'others' initiate repair. *Applied Linguistics 21*(2), 205–243.

Schlangen, D. (2004). Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*.

Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Pragmatics*, Volume 9 of *Syntax and Semantics*, pp. 315–332. New York Academic Press.