A recipe for annotating grounded clarifications

Luciana Benotti
Universidad Nacional de Córdoba
CONICET, Argentina
luciana.benotti@unc.edu.ar

Patrick Blackburn
Philosophy and Science Studies
IKH, Roskilde University, Denmark
patrickb@ruc.dk

Abstract

In order to interpret the *communicative intents* of an utterance, it needs to be *grounded* in something that is *outside of language*; that is, *grounded* in *world modalities*. In this paper we argue that dialogue clarification mechanisms make explicit the process of interpreting the communicative intents of the speaker's utterances by grounding them in the various modalities in which the dialogue is situated. This paper frames dialogue clarification mechanisms as an understudied research problem and a key missing piece in the giant jigsaw puzzle of natural language understanding. We discuss both the theoretical background and practical challenges posed by this problem, and propose a recipe for obtaining grounding annotations. We conclude by highlighting ethical issues that need to be addressed in future work.

1 Introduction

Clarifications are crucial to robust dialogues, and pragmatic factors — notably those shaped by the world modalities situating the conversation — have a key role to play. *Referring expressions* have in *vision* a modality in which to ground clarifications concerning objects in the world (de Vries et al., 2017); *navigation instructions* have in *movement* a modality in which to ground clarifications concerning collaborative wayfinding (Thomason et al., 2019). Clarifications grounded in situationally relevant modalities boost the *redundancy* required to learn to use language without explicit supervision, as they make explicit the process of negotiating the *communicative intent*. But despite its importance, work on clarification remains scattered.

Humans switch between clarifications grounded in different modalities seamlessly but (we shall argue) systematically. Our discussion is based around a general recipe for detecting *grounded clarifications*; we work towards this in Section 2 by first reviewing the distinction between perceptual and collaborative grounding, and then discussing clarification mechanisms, Clark (1996)'s action ladder of communication, and Ginzburg, Purver and colleagues (2012)'s classification of clarification phenomena. In Section 3 we draw these threads together and present the central idea:

> *Given an utterance U, a subsequent turn is its clarification grounded in modality* m *if it cannot be preceded by positive evidence of understanding of U in* m.

This provides a unified way to frame clarification mechanisms and their interactions across various modalities; a graphical specification of the recipe it gives rise to can be found in Figure 2 of the supplementary material. It covers clarifications grounded in moving, grabbing and changing the physical world: these have traditionally been considered plain-old-questions (Purver et al., 2018), but we view them as useful clarification ingredients.[1] In Sections 4 and A we test the practical implications of our recipe by identifying and characterizing (according to their modalities) the clarifications in a corpus of long dialogues in English. In Section 5 we turn to the claim that clarifications are rare in dialogue datasets (Ginzburg, 2012), and that current data-hungry algorithms cannot learn them. We argue that whether they are rare or not depends on pragmatic factors of the conversation and the modality of the grounded clarification, and discuss the impact of six such factors. After presenting potential objections and our responses in Section 6, we conclude in Section 7 by noting ethical issues raised by socioperceptive dialogue systems that will need to be addressed.[2]

[1] We are suspicious of the common assumption that requests for information regarding references that are grounded in vision (e.g. *the red or the blue jacket?*) are clarifications, whereas requests for information grounded in other modalities are not (e.g. *do I take the stairs up or down?*).

[2] See also the supplement on ethical considerations.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4065–4077
June 6–11, 2021. ©2021 Association for Computational Linguistics

2 Theoretical Background

We begin by reviewing the theoretical background on grounding and clarification mechanisms. We then examine two schemes proposed to characterize clarifications according to their conversational function: one focuses on the problem of anchoring utterance parameters into the conversational history, the other emphasizes a multimodal ladder of actions co-temporal with dialogue turn-taking. We are interested in the potential contributions of both towards a recipe for annotating clarification mechanisms.

2.1 Collaborative and perceptual grounding

Collaborative grounding is the process of seeking and providing incremental evidence of mutual understanding through dialogue. When the speaker believes that the dialogue is on track, *positive evidence* of understanding is provided in different forms (depending on the channel of communication) such as explicit acknowledgements, and via backchannels such as nods, eye contact, etc. *Negative evidence* of understanding signals that something needs negotiation before the dialogue partners can commit; clarification requests are the prototypical example of negative evidence.

Collaborative grounding is distinct from perceptual (or symbol) grounding (Harnad, 1990; He et al., 2016; Tan and Bansal, 2019; Lu et al., 2020). The perceptual grounding literature deals with capabilities enabling symbols to be linked with perceptions, and is rooted in situationally relevant modalities such as vision. Collaborative grounding, on the other hand, deals with the *dynamics of conversation* (the ongoing exchange of speaker and hearer roles) and is rooted in situationally relevant aspects of socioperception. Alikhani and Stone (2020) note several basic mechanisms that contribute to collaborative grounding, including those for dealing with joint attention (Koller et al., 2012; Koleva et al., 2015; Tan et al., 2020), engagement (Bohus and Horvitz, 2014; Foster et al., 2017), turn taking and incremental interpretation (Schlangen and Skantze, 2009; Selfridge et al., 2012; DeVault and Traum, 2013; Eshghi et al., 2015) corrections and clarifications (Villalba et al., 2017; Ginzburg and Fernández, 2010) and dialogue management (DeVault and Stone, 2009; Selfridge et al., 2012). These mechanisms have been studied for different kinds of applications (Denis, 2010; Dzikovska et al., 2010, 2012). Both collaborative and perceptual grounding are important (*all* relevant modalities are potentially important) and in this paper we bring them together under an umbrella we call *grounded clarification*.

2.2 Clarification mechanisms

Clarification requests (CRs) and their answers are the prototypical clarification mechanisms (CMs), pieces of dialogue that participants use to signal lack of understanding and to trigger negotiation. CMs are used in all kinds of dialogue and are influenced by the type of interaction, the dialogue participants, and the context in which the conversation occurs. Interest in CMs by the artificial intelligence community dates back to the start of the century, and has typically focused on mechanisms for human-computer dialogue systems (Gabsdil, 2003; Purver, 2004; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005; Skantze, 2007). In sociolinguistics and discourse analysis, on the other hand, the interest in CMs (or *repairs*, as they are usually called there) has focused on human-human conversation for over three decades now; see (Schegloff, 1987) for a representative example.

How CMs can be learned from data remains understudied. Rao and Daumé III (2018) rank clarification requests of stackoverflow articles according to their usefulness: a good clarification question is one whose expected answer will be useful, which means that the clarification highlighted important information missing from the initial request for help; we share this view, but differ from Rao and Daumé III, in that we focus on CMs and their responses occuring in multiturn dialogue.

It may seem plausible to expect that clarification requests will be realized as questions; however, corpus studies indicate that their most frequent realization is in declarative form (Jurafsky, 2004). Indeed, the form of a clarification request (Rodríguez and Schlangen, 2004) is *not* a reliable indicator of the function that the clarification request is playing. Neither does form unambiguously indicate whether a dialogue contribution is a CR or not. The surface form of explicit negotiations of meaning in dialogue are frequently non-sentential utterances (Fernández, 2006; Fernández et al., 2007). These include the prototypical positive and negative evidence of grounding (acknowledgements and clarification requests (Stoyanchev et al., 2013)) but also less-well-known forms such as self-corrections, rejections, and modifiers (Purver, 2004; Purver et al., 2018). These observations indicate that we face significant challenges if we want to train a system to seek or supply clarification effectively.

2.3 Clarifications grounded in parameters

Ginzburg, Purver and colleagues (henceforth G&P) proposed the first scheme to classify the functions of CRs; see (Purver et al., 2003; Purver, 2006; Ginzburg, 2012). The G&P classification uses the categories shown on Table 1. The idea driving this work is that CRs are caused by problems arising during the anchoring of utterance parameters into the conversational history.

CATEGORY	OBSTACLE	EXAMPLES
Repetition	Cannot identify a surface parameter	What did you say?
Clausal	Uncertain value for a clausal dialogue history parameter	Are you asking if BO SMITH left?
Intended	The hearer can find no value for a parameter	Who is Bo?
Correction	The hearer thinks that the speaker made a mistake and offers an alternative realization	Did you mean to say 'Bro'?

Table 1: CR classification scheme by P&G

The G&P classification has been criticized (Rodríguez and Schlangen, 2004; Rieser and Moore, 2005) because, in practice, it seems difficult to decide what the category of a particular CR is; that is, CRs are usually ambiguous in this classification. In fact, G&P recognize this issue themselves, pointing out that CRs that do not repeat (part of) the content of the *source utterance* (that is, the utterance that is being clarified) can exhibit all three readings.

However, G&P's classification is only ambiguous if *only the past, but not the future, conversational history is taken into account*. It is crucial to analyze the CR *response* in order to disambiguate the CR category. Sometimes the immediate linguistic context gives the clue necessary for disambiguation: whereas a repetition reading permits the responder to the CR to repeat her utterance verbatim, a clausal confirmation usually receives a yes/no answer, and an intended content reading requires the responder to reformulate in some way. Hence, the turn of the responder (and the subsequent reaction of the participant originally making the CR) can disambiguate among readings. Consider the following example from (Purver, 2004). The example shows a case where George's initial clausal interpretation is incorrect (the initiator is not satisfied), and a constituent reading is required (Anon cannot find a value for Spunyarn).

George: you always had er er say every foot he had with a piece of spunyarn in the wire
Anon: Spunyarn?
George: Spunyarn, yes
Anon: What's spunyarn?
George: Well that's like er tarred rope

In other situations, the immediate linguistic context will not be enough (for instance, a reformulation can be a good response to all three types of CRs) and then the whole conversational history might need to be analyzed in order to disambiguate. This makes G&P's classification difficult to use in annotation studies where the annotators only get shallow, partial, localized views of the dialogues.

2.4 Clarifications grounded in modalities

The second classification we shall examine puts the conversational action modality in the central role; it has been used in formal approaches to handling clarifications in dialogue systems (Gabsdil, 2003; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005). This classification is based on the four-level model of conversational action independently developed by (Allwood, 1995) and (Clark, 1996). Here, we use Clark's terminology; his model is reproduced in Table 2.

L	SPEAKER A'S ACTIONS	ADDRESSEE B'S ACTIONS
4	*Propose* project w to B	*Uptake* A's proposal w
3	*Intend* that B does i	*Recognize* i from A
2	*Present* signal s to B	*Perceive* signal s from A
1	*Execute* behavior t for B	*Attend* to behavior t from A

Table 2: Ladder of actions involved in communication

Clark proposed this model in order to move from Austin's controversial classification[3] of speech acts (Austin, 1962) to a *ladder of actions* which characterizes not only the actions that are performed in language use (as Austin's does) but also their inter-relationships. Clark (1996) defines a ladder of actions as a set of co-temporal actions which provide *upward causality* and *downward evidence*. Let us discuss these using Table 2; we will call the speaker Anna and the addressee Barny. Suppose that Anna tells Barny to sit down. We might say that Anna is performing just one action: asking

[3] For discussion of the controversies around Austin's classification of speech acts see (Clark, 1996)

Barny to sit down. But it is easy to argue that she is performing four distinct, though co-temporal, actions — actions beginning and ending simultaneously. These actions are in a *causal* relation going *up* the ladder (from level 1 up to level 4): Anna must get Barny to attend her behavior t (level 1) *in order to* get him to hear the words she is presenting in her signals (level 2). Anna must succeed at that *in order to* get Barny to recognize what she means (level 3), and she must succeed at that *in order to* get Barny to uptake the project she is proposing (level 4). In short, causality (do something *in order to* get some result) climbs up the ladder; this property Clark calls *upward causality*.

The different levels are related to different human modalities. We say that level 1 is grounded into *socioperception*, an ability that humans developed for collaboration that is crucial for achieving joint attention (Tomasello et al., 2005). Level 2 is grounded in *hearing* if we use speech as our communication channel. Level 3 is grounded in *vision* when it involves recognizing referents in the real world. Level 4 is grounded in *kinesthetic* when it involves moving and acting in the real world. The classification, along with obstacles that the addressee may face in the various modalities during the interpretation of a conversational action, is shown in Table 3. In the rest of the paper we will refer to these modalities using the level number.

L	MODALITY	EXAMPLES
4	Kinesthetic	Do I take the stairs up or down?
3	Vision	The red or the blue jacket?
2	Hearing	What did you say?
1	Socioperception	Are you talking to me?

Table 3: Ladder of actions grounded in modalities.

Humans systematically use the evidence provided by this ladder. Observing Barny sitting down is good evidence that he did not refuse to uptake (level 4) but also recognized what Anna intended and identified the chair (level 3). That is also evidence that she got Barny to hear her words (level 2), and evidence that she got him to attend to her (level 1). That is, evidence trickles *down* the ladder; Clark calls this the *downward evidence* property.

If Barny repeats verbatim what Anna said (e.g. suppose she spoke in Spanish and he repeats the word *sientate*), then Anna has good evidence that he heard what she said (level 2). However, that is not necessarily evidence that he has recognized her intention; there might be an obstacle in level 3 (for instance, Barny might not know Spanish). If there is such an obstacle, she would have completed levels 1 and 2 while failing to complete not only level 3 but also level 4 (it is rather unlikely that Barny would sit down right after hearing Anna — and even if he did, this would not be *because* he was uptaking Anna's project). A high level action in the ladder can only be completed by executing *all* the actions in the lower levels. This property Clark calls *upward completion*.

If you tell somebody something, you expect a reaction from him. If he doesn't answer, you might think that he didn't hear you, that he doesn't want to answer, or that he thinks you are talking to somebody else. None of these situations is very agreeable; humans don't like wasting effort, or being ignored. In order not to annoy the speaker, the addressee has two options: either he shows evidence in level 4 (and then, by downward evidence, the speaker knows that all the levels succeeded), or he indicates the obstacle in executing the action (in any level). Clarifications are the tools that addressees can use to make the obstacle explicit.

3 A grounded clarification recipe

In this section we draw these threads together under the heading *grounded clarification*. First, what is a clarification? Our starting proposal, which we will modify, is the following: *given an utterance U, a subsequent turn is its clarification if it cannot be preceded by positive evidence of U*. Note that this proposal implicitly embodies a procedure for annotating clarifications, one which could be crowdsourced: *Is this a clarification? Check whether it can be preceded by positive evidence!*

Our starting proposal is a modified version of Gabsdil (2003)'s test for CRs. Gabsdil says that CRs (as opposed to other kinds of dialogue contributions) cannot be preceded by explicit acknowledgments. For example:

Lara: There's only two people in the class.
a) Matthew: Two people?
b) (*) Matthew: Ok, Two people?
(BNC, taken from (Purver et al., 2003))

Gabsdil argues that (a) in the example above *is* a CR because (b) is odd (we mark odd turns with (*) in examples). In (b), Matthew first acknowledges Lara's turn and only then indicates that her turn contains information that he finds controversial.[4]

[4] This could be a felicitous response, but it would require

On the other hand, (b) in the example below is fine and hence (a) *is not* a CR: the lieutenant acknowledges the sergeant's turn and then moves on to address what has become the most pressing topic in the conversation:

Sergeant: There was an accident sir
a) Lieutenant: Who is hurt?
b) Lieutenant: Ok. Who is hurt?
Adapted from (Traum, 2003, p.391)

However Gabsdil's original test incorrectly discards cases that we view as CRs. Consider the following example:

G: I want you to go up the left hand side of it towards the green bay and make it a slightly diagonal line, towards, sloping to the right.
F: Ok. So you want me to go above the carpenter?
Adapted from (Gabsdil, 2003, p.30)

The problem is that the level of positive evidence contributed by F's acknowledgment is ambiguous. For instance, the Ok could (conceivably) mean:

- *Ok, so you want to talk to me* (level 1).
- *Ok, I heard you* (level 2).
- *Ok, I saw what you are referring to* (level 3).
- *Ok, I did it* (level 4, the highest level).

Thus we modify Gabsdil's test to make it *level-sensitive*. In order to signal that all the levels have been successful and that no CR related to any of them is expected, the simple acknowledgment needs to be replaced by positive evidence in the highest level. This works for Gabsdil's example:

G: I want you to go up the left hand side of it towards the green bay and make it a slightly diagonal line, towards, sloping to the right.
(*) F: Ok, I did it. So you want me to go above the carpenter?

Here *So you want me to go above the carpenter?* is either weird or far more likely to be interpreted as a question about an action that comes *after* F has successfully followed G's instruction. That is: it could be interpreted as F taking the initiative and proposing the next move, rather than as clarifying G's instruction. Whether this is plausible would be determined by the following turns.

More generally, if the addressee wants to uptake the speaker's proposal then he or she has two options: either to give positive evidence at the highest modality (and then, by downward closure, the speaker knows that all lower levels succeeded) or to explicitly indicate the problem using a clarification (at any level). Table 3 illustrates, for each level and modality, possible CRs. We are not exhaustive about all the modalities that could happen in reality. We list four of them here but there could be more depending on the task.

This approach to CR identification and classification is useful not only for instructions but also for other types of utterances. The following is an extension of Grice's classic implicature example (physical actions are between square brackets):

A: I am out of petrol.
B: There is a garage around the corner.
A: [A goes to the garage and then meets B again]
(*) A: Ok, I got petrol at the garage. Do you think the garage was open?
Adapted from (Grice, 1975, p.311)

After acknowledging a contribution at level 4 (which A's *Ok, I got petrol at the garage* clearly does) it is really hard to go on and ask a CR about that contribution (A's *Do you think the garage was open?* is a bizarre follow-up — it could perhaps be interpreted as sarcastic).

Thus our modified proposal for identifying clarifications is the following: *given an utterance U, a subsequent turn is its clarification grounded in modality* m *if it cannot be preceded by positive evidence of understanding of U in* m.[5] Like the earlier version, this implicitly embodies a annotation procedure. Let's see how it works.

4 Grounded clarification annotation

In this section we evaluate our recipe and the modality-based classification it gives rise to. We do so by using it to annotate a small dataset, the SCARE corpus (Stoia, 2007). Before delving into the details of the classification, we describe the pragmatic influences that the dialogue participants are under in this dataset.

The SCARE corpus consists of fifteen English spontaneous dialogues situated in an *instruction giving task*.[6] The dialogues vary in length, with a

marked intonation to induce a backtracking effect.

[5] For a detailed graphical specification of our recipe, see Figure 2 in the supplementary material. Notice that the utterances are stored in a stack in Figure 2 because the clarification does not need to be immediately after its source. While an utterance is at the top of the stack it can be clarified, no matter how many turns in between have happened. That way an utterance can be clarified many times.

[6] The corpus is available at http://slate.cse.ohio-state.edu/quake-corpora/scare/.

minimum of 400 turns and a maximum of 1500; hence, the dialogues are *much* longer than other datasets grounded in vision and action where dialogues typically have less than 10 turns on average (de Vries et al., 2017; Thomason et al., 2019). The dialogues were collected using the QUAKE environment, a first-person virtual reality game (so there is *immediate world validation*). The task consists of a direction giver (DG) instructing a direction follower (DF) on how to complete several tasks in a simulated game world. The corpus contains the collected audio and video, as well as word-aligned transcriptions.

The DF had no prior knowledge of the world map or tasks and relied on his partner, the DG, to guide him on completing the tasks (so the DPs have *asymmetric knowledge of the task*). The DG had a map of the world and a list of tasks to complete. The partners spoke to each other through headset microphones. As the participants collaborated on the tasks, the DG had instant feedback about the DF's location in the simulated world, because the game engine displayed the DF's first person view of the world on both the DG's and DF's computer monitors (so the DPs *share a view of the task*). Finally, the DPs were *punished* (they were told they would receive less money for performing the experiment) if they pressed the wrong buttons or put things in the wrong cabinets.

We present a sample interaction from the SCARE corpus. During this dialogue fragment, the dialogue participants were performing one of the tasks of the SCARE experiment specified: *hide the rebreather in cabinet 9*.

The presentation of this dialogue is divided over the two following subsections; the first gives the warm-up necessary for the second. Subsection 4.1 illustrates how positive evidence of understanding is provided, and no examples of CRs are presented here. Subsection 4.2's goal, on the other hand, is to illustrate CRs in different modalities, so here we focus on negative evidence.

4.1 Positive evidence

At the beginning of this dialogue, the DG is instructing the DF to find the rebreather. As part of this task, they have to press a button in order to open a door as shown in Figure 1. The figure shows a dialogue fragment and a screenshot of the shared view when the fragment starts. The turns which provide positive evidence at levels 3 and 4 are shown in boldface. If evidence for proposal is followed by a turn that is not evidence of uptake (of the proposal) then we say that the turn is a CR.

DG(1): see that button straight ahead of you?
DF(2): ***mhm***
DG(3): hit that one
DF(4): ***ok***

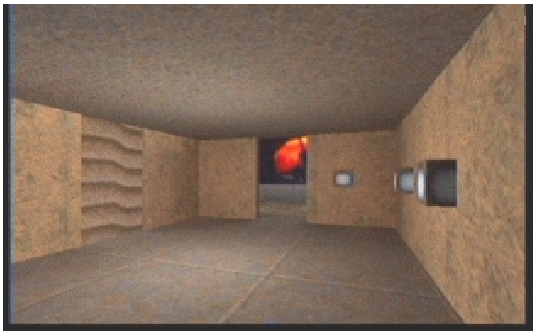

Figure 1: Example of the view shared by the dialogue participants and fragment from the SCARE corpus

The dialogue fragment reproduced below starts when the DG is trying to get the DF to press the button that is straight ahead in their current view; this button opens the cabinet where the rebreather is located. As part of this project, the DG first makes sure that the DF identifies this button using the sub-dialogue constituted by (1) and (2). Once the button is identified, the short instruction in (3) suffices to convey the goal of the joint project, namely hitting this button; this is acknowledged at level 4 in turn (4) when the DF presses the button.

4.2 Negative evidence

Now we turn to an extended example, extracted from the SCARE corpus, of clarification requests at different levels. Between square brackets we indicate forms of non-linguistic communication. The DG utters an instruction in (1). In turn (2) the DF makes explicit an obstacle at level 3 that must be solved *before* putting the rebreather in the cabinet, namely identifying cabinet 9; in doing so he proposes this task. In turn (3) the DG proposes to identify cabinet 9 by first identifying its location. Turn (4) is evidence of uptake of turn (3) — the DG answers his own question — but it is also evidence of the proposal: get back to the starting room.

DG(1): we have to put it in cabinet nine [pause]
DF(2): yeah they're not numbered [laughs]
DG(3): [laughs] where is cabinet nine
DG(4): it's kinda like back where you started so
DF(5): ok so I have to go back through here?

DG(6): yeah
DF(7): and around the corner?
DG(8): right
DF(9): and then do I have to go back up the steps?
DG(10): yeah
DF(11): alright this is where we started
DG(12): ok so your left ca-[pause] the left one
DF(13): so how do I open it?
DF(14): one of the buttons?
DG(15): yeah, it's the left one
DF(16): makes sense
DF(17): alright so we put it in cabinet nine

Of the 17 turns, 9 were uttered by the DF and 8 by the DG. From the 9 turns by the DF, 5 of them are CRs at level 4 and one at level 3. Turn (2) is a CR of instruction (1). Turns (5), (7) and (9) are CRs of instruction (4). Utterance (11) shows positive evidence at level 4 of instruction (4) so this instruction cannot be further clarified following the recipe we defined in Section 3. Turns (13) and (14) are CRs of utterance (12). The positive evidence at level 4 of instruction (12) is completed by a physical action of the DF in the game world: opening the cabinet by pressing the left button while uttering (16). Finally, turn (17) together with the corresponding physical action are positive evidence at level 4 of instruction (1).

5 Comparative analysis of clarifications

In this section, we identify and discuss a number of pressures that interact in order to determine the number and type of CRs that occur in dialogue; we also explain why it makes sense (although it may seem counter-intuitive at first sight) that too much uncertainty will tend to *lower* the number of CRs.

The distribution and types of CRs found in a corpus depend on the characteristics of the task that the dialogues in the corpus are addressing. Previous clarification corpus studies (Purver, 2004; Rieser and Moore, 2005; Rodríguez and Schlangen, 2004) have required expensive and detailed annotations by linguists who also evaluated the quality of the datasets. Purver (2004) annotates more than 10K turns of the BNC corpus, which contains English dialogue transcriptions of topics of *general interest* in multiparty dialogue such as meetings. These annotations were used to build a dialogue system that could make and understand relevant clarifications related to different modalities (Purver, 2006). (Rieser and Moore, 2005) and (Rodríguez and Schlangen, 2004) did similar annotations on task-oriented dialogue corpora. (Rieser and Moore, 2005) looked for CRs in a corpus of English task-oriented human-human dialogue called Communicator. The corpus consists of travel reservation dialogues between a client a travel agent. The interactions occur by phone; the participants do not have a shared view of the task. The corpus comprises 31 dialogues of 67 turns each (on average), from which 4.6% of the turns are CRs. 12% of CRs found were classified as level 4 CRs; such as the following:

> *Client: You know what the conference might be downtown Seattle so I may have to call you back on that.*
> *Agent: Okay. Did you want me to wait for the hotel then?*

In this corpus the world validation is informational not physical as in the Bielefeld data that we turn to now.

(Rodríguez and Schlangen, 2004) looked for CRs in a corpus of German task-oriented human-human dialogue called Bielefeld. The dialogues occur in a instruction giving task for building a model plane. The interactions occur face to face; the participants have a shared-view of the task. The corpus consists of 22 dialogues, with 180 turns each (on average), from which 5.8% of the turns are CRs. 22% of CRs found were classified as level 4 CRs, such as the following:

> *DG: Turn it on.*
> *DF: By pushing the red button?*

We analyzed the SCARE corpus while watching the associated videos and we classified the clarification requests according to the levels of communication using the decision procedure explained in Section 3.[7] We found that 6.5% of the turns are CRs. Of these, 65% belong to level 4 of Table 2, and 31% belong to level 3 (most of them related to reference resolution). Only 2% of the CRs were acoustic (level 2) since the channel used was very reliable, and another 2% had to do with establishing contact (level 1).

The SCARE corpus presents slightly more CRs (at 6.5%) than the corpora analyzed in previous work (which reported that 4%-6% of the dialogue turns were CRs). Furthermore, in contrast to the

[7] We will release our annotations to the research community upon request.

BNC corpus study (Purver, 2004), most CRs in the SCARE corpus occurred at level 4. What task characteristics might have caused the observed differences?

We hypothesize that the following six characteristics account for the larger proportion of CRs at level 4 that we find in the SCARE corpus. *Task oriented* dialogues (unlike general interest dialogues) are constrained by the task, thus the hearer may have a better hypothesis of what the problem is with the source utterance. He also has a clear motivation for asking for clarifications when the utterance does not fit his model of the task. Dialogues situated in an instruction giving task show an *asymmetry* between the knowledge that the dialogue participants (DPs) have about the task. The Direction Giver (DG) knows how the task has to be done and the Direction Follower (DF) doesn't. Hence, it is to be expected that the DF will have doubts about the task which (both DPs know) can only be answered by the DG. In symmetric dialogues, it might not be clear who has what information and then the DPs might not know who can answer the CRs. *Immediate world validation* seems to play a role as well. Dialogues that interleave linguistic actions and informational or physical actions exhibit immediate world validation of the interpretations. If an instruction fails in the world, the DF will ask for clarification. When the DPs have a *shared view* of the task, the DP that is acting on the world knows that the other participant is observing him and verifying his actions and then will try to be sure of what he has to do before doing it. If he is not sure he will ask. *Long dialogues* tend to increase the percentage of clarifications (more than 100 turns) because DPs prefer to ask questions when they have a good hypothesis to offer. The longer the interaction, the more background is shared by the DPs and the easier it will be to come up with a good hypothesis. Finally, if there are actions in some modality that are *irreversible*, then they will clarify more until they are sure of what they have to do.

6 Discussion and objections

Humans switch between clarifications grounded in different modalities seamlessly and we have argued they do so systematically; in effect they do so by following a recipe for grounding classifications. We obtained this recipe by granting a role to both perceptual and collaborative grounding in clarification requests. This we did by examining Clark (1996)'s action ladder of communication and Ginzburg, Purver and colleagues (2012)'s classification of clarification phenomena, and combining the concept of level taken from the ladder of communication with Gabsdil (2003)'s test for clarification requests. We reframed Clark's downward evidence and upwards completion properties for multimodal interactions.

This gave us the following: *given an utterance, a subsequent turn is its clarification grounded in modality* m *if it cannot be preceded by positive evidence of understanding in* m. This provides a unified way to frame clarification mechanisms and their interactions across modalities — something we view as useful in its own right given the scattered literature on clarification mechanisms. However we also suggested that this recipe was suitable for learning from data collected by crowdsourcing. We supported this by examining the claim that clarifications are rare in dialogue datasets (Ginzburg, 2012), and that current data-hungry algorithms cannot learn them. We argued that whether they are rare or not depends on pragmatic factors of the conversation and the modality of the grounded clarification. Moreover, along the way we noted a number of practical issues — work with *large* dialogues, don't just provide annotators with dialogue fragments, take future conversational history into account when annotating — that we think could have an important impact on learnability.

Below we list some possible objections to our proposal. We also include our responses in the hope that this will motivate further debate on these issues in the community.

Objection: I still don't have a feel for how much we will gain from this when it comes to a practical, realistic use case; in particular, for an end-to-end system rather than an NLP pipeline.

Response: Being able to identify and annotate a turn as a clarification request can help an end-to-end system learn to apply the mechanisms of collaborative grounding to subdialogs, which have rules that differ from modality to modality.

Objection: The biggest problem I see is that the distinction of the different levels (which the correct annotation relies on) might not be clear-cut (in particular when considering that crowdsourced annotations usually come from non-experts). I have no idea what quality we get, nor what inter-annotator agreement figures we can expect.

Response: Our methodology unifies and refines

Characteristics	BNC fragment	Communicator	Bielefeld	Scare
Task	Chit-chat	Travel reservation	Building	Moving
Shared view	Yes (meetings)	No (on the phone)	Yes (face-to-face)	Yes (3D game)
Participants	More than two	Two	Two	Two
World validation	Common ground	Informational	Physical	Simulated
Information Flow	Symmetrical	Symmetrical	Symmetrical	Asymmetrical
Total # turns	10466	2098	3962	11350
Avg dialogue length	30	67	180	800
% of CRs/turns	3.1	4.6	5.8	6.8
% CRs level 1	10	3	0	3
% CRs level 2	31	32	12	9
% CRs level 3	47	40	50	32
% CRs level 4	2	12	22	53
% CRs other	10	13	16	4

Table 4: Comparing the number of CRs at each level in four corpus studies

previous methodologies for which inter-annotator agreement has been reported in certain corpora: E.g., .70 for the Bielefeld corpus (Rodríguez and Schlangen, 2004), and .75 for the BNC corpus (Purver, 2004). Our methodology refines Clark (1996) 4-level classification by grounding each level (previously only described by means of examples) to 4 different modalities relevant for situated dialog: socioperception, hearing, vision and movement. This new grounded characterization should improve previous inter-annotator agreement. Using our extended methodology we report a .84 kappa for the SCARE.

Objection: The corpora that are being investigated are all very domain-specific and relatively small in terms of numbers of dialogues (but with a large average number of turns). This means that even if we were to obtain annotation quality figures, it would still raise the question of what general conclusions we can draw from this.

Response: We share this concern; our goal with this paper is to motivate more work in this area. We believe that this objection actually lends support to our insistence on the importance of a more fine-grained analysis of grounding mechanisms. Our methodology generalizes to domains that ground the communicative intent in the modalities of socioperception, hearing, vision and movement. Examples are robots and virtual assistants, where the dialog partners share a sensible environment. Our argument is that better conceptualizations of clarification subdialogs are needed so that models are able to identify them, distinguish the different types ruled by the different modalities, and learn the structures that govern them.

7 Conclusions

This paper urges the community to address a research gap: how clarification mechanisms can be learned from data. We believe that novel research methodologies which highlight the importance of the role of clarification mechanisms in communicative intent are needed for this. So we presented an annotation methodology, based on a theoretical analysis of clarification requests, which unified a number of previous accounts.

But to conclude, a different note. As dialogue systems get better at negotiating meaning with clarifications, future work will need to seriously consider how people relate to conversationally-gifted artificial agents. Studies of how users feel when interacting with dialogue systems (Brave et al., 2005; Portela and Granell-Canut, 2017) found that systems can have a psychological impact on users; thus it will become increasingly important to consider the risks of users developing social or emotional bonds with more sophisticated system (thereby affecting their well-being in unforseen ways) and of users being emotionally manipulated by them. Socioperceptive dialogue systems could turn out to have very sharp teeth indeed.

Acknowledgments

We thank the anonymous reviewers for their detailed reviews and insightful comments.

Ethical considerations

In this paper we have not trained machine learning models so we have used negligible computing power. We have not collected a new dataset so we have not used crowdsourcing. The annotation of the SCARE corpus was done by one of the authors and a friend who likes the work and was not economically rewarded. As we noted in the papers conclusion, there are important ethical issues that future work on this area will need to consider. But there are also more immediate discuss ethical considerations and we turn to these now.

First, the datasets that we use in this paper are described in (Purver, 2004; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005; Stoia, 2007). The dataset in (Purver, 2004) contains spoken British English dialogues collected during meetings. The dataset used in (Rieser and Moore, 2005) is a fragment of the Carnegie Mellon Communicator Corpus (Bennett and Rudnicky, 2002), and is in American English. In these dialogues, an experienced travel agent is making reservations for trips that people in the Carnegie Mellon Speech Group were taking in the upcoming months. There is no information to whether the dialogue participants were rewarded or notified about the dataset collection. The dataset in (Rodríguez and Schlangen, 2004) includes dialogues in which one participant gives instructions in German to the other to build a model plane. Finally, the Scare corpus (Stoia, 2007) is an American English corpus collected using students at Ohio State University; they were payed to participate in the experiment.

Future work in this area will need to collect new datasets that reflect the interactions between different types of clarifications in different modalities. Usually such collections are crowdsourced, which raises ethical concerning fair wages and number of hits per day. We would like to encourage the community to value datasets in languages other than English in order to model different strategies for indicating the source of the clarification (prosody, syntactic construction, etc). Last but not least, computing power and carbon footprint should be considered. Machine learning models trained on long multimodal dialogue histories may get very big very fast. We need models that learn to summarize dialogue histories for the sake of the environment and the budget of low-income researchers.

References

Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15. ACL.

Jens Allwood. 1995. An activity based approach to pragmatics. In Harry Bunt and William Black, editors, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, pages 47–80. John Benjamins Publishers.

John Austin. 1962. *How to do Things with Words: The William James Lectures delivered at Harvard University in 1955.* Oxford University Press.

Christina L. Bennett and Alexander I. Rudnicky. 2002. The carnegie mellon communicator corpus. In *Interspeech*. ISCA.

Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 2–9, New York, NY, USA. ACM.

Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2):161–178.

Herbert Clark. 1996. *Using Language*. Cambridge University Press, New York.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.

Alexandre Denis. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the Sixth International Natural Language Generation Conference*. ACL.

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 184–192. ACL.

David DeVault and David Traum. 2013. Approximation of incremental understanding of explicit utterance meaning using predictive models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1092–1099. ACL.

Myroslava Dzikovska, Peter Bell, Amy Isard, and Johanna Moore. 2012. Evaluating language understanding accuracy with respect to objective outcomes in a dialog system. In *Proceedings of the Conference European Chapter of the Association for Computational Linguistics*, pages 471–481. ACL.

Myroslava Dzikovska, Johanna Moore, Natalie Steinhauser, and Gwendolyn Campbell. 2010. The impact of interpretation problems on tutorial dialogue. In *Proceedings of the Conference of the Association for Computational Linguistics*, pages 43–48. ACL.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271, London, UK. ACL.

Raquel Fernández. 2006. *Non-sentential utterances in dialogue: Classification, resolution and use*. Ph.D. thesis, University of London.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2017. Automatically classifying user engagement for dynamic multi-party human-robot interaction. *International Journal of Social Robotics*, 9(5):659–674.

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Bart Geurts. in press. *Quantity implicatures*. Cambridge University Press.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford Press.

Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Blackwell.

Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, New York.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Daniel Jurafsky. 2004. Pragmatics and computational linguistics. In *Handbook of Pragmatics*, pages 3–28. Blackwell, Oxford.

Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. 2015. The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 812–817. ACL.

Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 30–39. ACL.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10434–10443. IEEE.

Manuel Portela and Carlos Granell-Canut. 2017. A new friend in our smartphone?: observing interactions with chatbots in the search of emotional engagement. In *Proceedings of the XVIII International Conference on Human Computer Interaction, Interacción*, pages 48:1–48:7. ACM.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London. Supervised by Jonathan Ginzburg.

Matthew Purver. 2006. CLARIE: Handling clarification requests in a dialogue system. *Research on Language and Computation*, 4(2-3):259–288.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and New Directions in Discourse and Dialogue*, pages 235–255. Kluwer Academic Publishers.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proc of ACL*, pages 239–246.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task oriented spoken dialogues. In *Proc of SEMDIAL*, pages 101–108.

Emanuel Schegloff. 1987. Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 8:201–218.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 710–718. ACL.

Ethan O. Selfridge, Iker Arizmendi, Peter A. Heeman, and Jason D. Williams. 2012. Integrating incremental speech recognition and POMDP-based dialogue systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–279, Seoul, South Korea. ACL.

Gabriel Skantze. 2007. *Error Handling in Spoken Dialogue Systems*. Ph.D. thesis, KTH - Royal Institute of Technology, Sweden. Supervised by R. Carlson.

Laura Stoia. 2007. *Noun Phrase Generation for Situated Dialogs*. Ph.D. thesis, Ohio State University, USA. Supervised by Donna Byron.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. In *Proceedings of the SIGDIAL 2013 Conference*, pages 137–141, Metz, France. ACL.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL.

Xiang Zhi Tan, Sean Andrist, Dan Bohus, and Eric Horvitz. 2020. Now, over here: Leveraging extended attentional capabilities in human-robot interaction. In *Companion of the International Conference on Human-Robot Interaction*, HRI '20, page 468–470, New York, NY, USA. ACM.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *3rd Annual Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 394–406. PMLR.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proceedings of the International Workshop on Computational Semantics (IWCS)*, pages 380–394.

Martín Villalba, Christoph Teichmann, and Alexander Koller. 2017. Generating contrastive referring expressions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Vancouver, Canada. ACL.

A Annotation decision procedure

In this section we formalize the classification methodology for clarifications. Our examples here are from the SCARE corpus (Stoia, 2007). The SCARE corpus consists of fifteen English spontaneous dialogues situated in an *instruction giving task*.[8] The dialogues vary in length, with a minimum of 400 turns and a maximum of 1500.

The annotation was performed by two independent annotators with an initial interannotator agreement of .84 kappa. Disagreements were discussed until agreement, a single annotation was obtained. We excluded from the annotation dialogue 1 which has almost no feedback from the DF because he thought that he was not supposed to speak.

The decision graph used in our "quasi-systematic" annotation is depicted in Figure 2. We call our procedure "quasi-systematic" because, while its tasks (depicted in rectangles) are readily automated, its decision points are not as they require subjective human judgments. Decision points D1 and D2 decide whether the turn is a CR or not; new tasks and digressions from the current task answer "no" to both decision points and just stack their evidence of proposal in T3. If the turn is a CR of a proposal X, T4 unstacks all proposals over X as a result of applying the downward evidence property of conversations (discussed in the paper). Intuitively, the turn is taken as an implicit uptake in level 4 of all the proposals over proposal X (which must be completed before X can be completed).[9] Decision points D3 to D6 decide whether the CRs belong to (Clark, 1996)'s levels 1 to 4 respectively, with the help of (Rodríguez and Schlangen, 2004). If a turn can be preceded by positive evidence in level 4 but it still is negative evidence of some proposal made earlier in the dialogue we annotate the CR as *other*. An example of this in dialogue 2 in the SCARE dataset *THERE'S ISN'T REALLY SHORT FOR THERE ARE, IS IT? BUT PEOPLE DO IT ANYWAY* where the DF follower is correcting the DG who said *THERE'S THREE DOORS* earlier. The negative evidence is not related to the modalities relevant for the task at hand but to the language itself.

[8] The corpus is available in http://slate.cse.ohio-state.edu/quake-corpora/scare/.

[9] This intuition is in line with Geurts's preliminary analysis of non-declaratives: if the speaker did not negotiate the proposals over X, then we can assume that he did not have problems up-taking them (Geurts, in press).

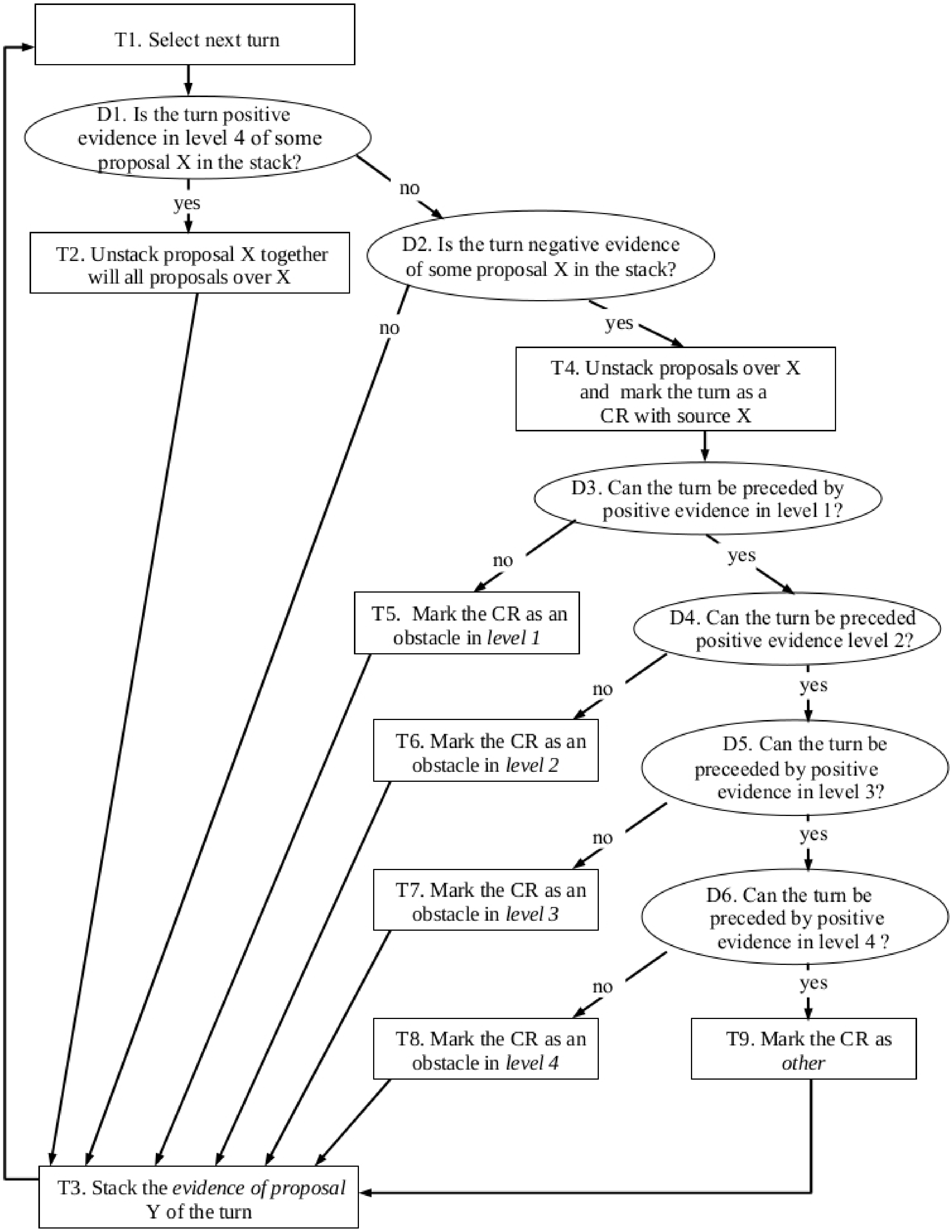
T1. Select next turn
D1. Is the turn positive evidence in level 4 of some proposal X in the stack?
yes
no
T2. Unstack proposal X together will all proposals over X
D2. Is the turn negative evidence of some proposal X in the stack?
no
yes
T4. Unstack proposals over X and mark the turn as a CR with source X
D3. Can the turn be preceded by positive evidence in level 1?
no
yes
T5. Mark the CR as an obstacle in *level 1*
D4. Can the turn be preceded positive evidence level 2?
no
yes
T6. Mark the CR as an obstacle in *level 2*
D5. Can the turn be preceeded by positive evidence in level 3?
no
yes
T7. Mark the CR as an obstacle in *level 3*
D6. Can the turn be preceded by positive evidence in level 4 ?
no
yes
T8. Mark the CR as an obstacle in *level 4*
T9. Mark the CR as *other*
T3. Stack the *evidence of proposal* Y of the turn

Figure 2: Decision graph for our recipe. Decision points D1 and D2 decide whether the turn is a CR or not. Decision points D3 to D6 decide whether the CRs belong to (Clark, 1996)'s levels 1 to 4 respectively. T9 indicates that the CR is grounded in a modality not represented by levels 1 to 4.