INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

ILLC PhD Pilot Study

Working Title	Agreement and Disagreement in Dialogue
Name PhD Candidate	Julian J. Schlöder
Name Supervisors	Raquel Fernández (main supervisor), Robert A. M. van Rooij (promotor).
Date	December 15, 2014.
Report	As a pilot study I have appended two papers I submitted to the International Conference on Computational Semantics (IWCS '15) on December 15th, 2014. The first paper is on the topic of *pragmatic rejection*, where I describe a class of utterances with rejection function, that (a) are consistent with the utterance they are rejecting and (b) cannot be circumscribed as asserting the negation of their antecedent. Both (a) and (b) are in contrast with previous knowledge. I discuss a simple formal model, a simulation of that model and an empirical experiment. The second paper describes a corpus study on the notion of *uptake-level* (or intention-level) *clarification requests*. It describes the data I have gathered to continue working on the notion of *agreement in dialogue* by way of investigating clarification requests that go beyond semantic parsing.

Work Plan

I have separated my project in two subprojects: Uptake and Clarification, and Negation and Rejection. By setting the plans up (more or less) in parallel, I can maintain the two lines of research simultaneously, but need to familiarise myself with only one set of methods at a time.

Uptake and Clarification

- A **conceptual model** (based on my Master's thesis) already exists.
- Based on the AMI corpus, I have assembled a **corpus of (possible) Uptake Clarifications**.
- A **pilot study** of the corpus is presented in the paper on uptake.
- A **full corpus study** will be completed by early-mid 2015.
- In parallel, at my Edinburgh secondment a **formalisation** will be put down, and refined afterwards (by late 2015).
- The formal model will be **evaluated** from a philosophical-conceptual standpoint.
- As soon as a formal framework is established, the theory will be **expanded** to cover more data.

Negation and Rejection

- I have gathered **observations** on different types of rejections in natural dialogue.
- A **toy model** to explain implicature rejections exists including a **framework for simulation**; presented in the paper on pragmatic rejection.
- In 2015, this model will be expanded to a full **formal model**.
- I have **assembled a corpus** of interesting rejections from different sources.
- I am conducting a broader philosophical investigation into the problem of **whether or not rejection is negative assertion**.
- Based on this investigation and the formal model for agreement (above), I will put down a **model for rejection**.

This timeline should be considered tentative. I expect further details and research questions to emerge as my research develops further.

Pragmatic Rejection

Abstract

Computationally detecting the accepting/rejecting force of an utterance in dialogue is often a complex process. In this paper we focus on a class of utterances we call *pragmatic rejections*, whose rejection force arises only by pragmatic means. We define the class of pragmatic rejections, present a novel corpus of such utterances, and introduce a formal model to compute what we call *rejections-by-implicature*. To investigate the perceived rejection force of pragmatic rejections, we conduct a crowdsourcing experiment and compare the experimental results to a computational simulation of our model. Our results indicate that models of rejection should capture partial rejection force.

1 Introduction

Analysing meaning in dialogue faces many particular challenges. A fundamental one is to keep track of the information the conversing interlocutors mutually take for granted, their *common ground* (Stalnaker, 1978). Knowledge of what is—and what is not—common ground can be necessary to interpret elliptical, anaphoric, fragmented and otherwise non-sentential expressions (Ginzburg, 2012). Establishing and maintaining common ground is a complicated process, even for human interlocutors (Clark, 1996). A basic issue is to determine which proposals in the dialogue have been *accepted* and which have been *rejected*: Accepted proposals are committed to common ground; rejected ones are not (Stalnaker, 1978). An important area of application is the automated summarisation of meeting transcripts, where it is vital to retrieve only mutually agreed propositions (Galley et al., 2004).

Determining the acceptance or rejection function of an utterance can be a highly nontrivial matter (Walker, 1996; Lascarides and Asher, 2009) as the utterance's surface form alone is oftentimes not explicit enough (Horn, 1989; Schlöder and Fernández, 2014). Acceptance may merely be inferable from a *relevant next contribution* (Clark, 1996), and some rejections require substantial contextual awareness and inference capabilities to be detected—for example, when the intuitive meaning of *'yes'* and *'no'* is *reversed*, as in (1), or when the rejection requires some *pragmatic enrichment*, such as computing presuppositions in (2):[1]

(1) A: TVs aren't capable of sending.
B: Yes they are.
$\rightsquigarrow$ *rejection*

(2) A: You can reply to the same message.
B: I haven't got [the] message.
$\rightsquigarrow$ *presupposition failure*

Our main concern in this paper are rejections like (2) whose rejection force can only be detected by pragmatic means. Aside from presupposition failures, we are particularly concerned with rejections related to implicatures: either rejections-of-implicatures or rejections-by-implicature as in the following examples of scalar implicatures:[2]

(3) A: That's brilliant.
B: Well I thought that was quite good.
$\rightsquigarrow$ *good, not necessarily brilliant*

(4) A: It was good weren't it?
B: It's brilliant.
$\rightsquigarrow$ *not merely good*

In both examples, B's utterances do not seem to (fully) agree with their antecedent: In (3) B can be taken to implicate *'good* $\rightsquigarrow$ *not brilliant'*, thereby disagreeing with A's assertion; in (4), B can be taken

[1] Examples from the AMI Meeting Corpus (Carletta, 2007).

[2] Examples from the British National Corpus (BNC) (Burnard, 2000).

to reject the same implicature attributed to A. We consider both examples to be what we call *pragmatic rejections:* utterances whose rejection force is indeterminable by purely semantic means. A particular feature of such rejections is that they are *prima facie* not in logical contradiction with their antecedent. Yet, as pointed out by Walker (2012), a widespread consent identifies rejection force with contradicting content.

We proceed as follows: In the next section, we give a more comprehensive account of what we introduced in the previous paragraph, offer a precise definition of the term *pragmatic rejection*, and discuss some associated problems. Afterwards, we review related literature, both on the topic of rejection computing and on the pragmatics of positive and negative answers. The main contributions of our work are a novel corpus of pragmatic rejections (Section 4), a formal model to compute rejections-by-implicature (Section 5), and a crowdsourcing experiment to gather agreement/disagreement judgements. In Section 6, we present the results of this experiment and compare them to a computational simulation of our model. We summarise our findings and conclude in Section 7.

2 Pragmatic Rejection

A commonly held view on rejection states that a speech event constitutes a rejecting act if and only if it is inconsistent in the dialogue context (*e.g.*, in the formal model of Schlöder and Fernández, 2014). Under that conception, rejection is typically modelled as asserting the negative of a contextually salient proposition. However, as observed by Walker (1996, 2012), this does not give the full picture. A perfectly consistent utterance can have rejection force by a variety of *implicated* inconsistencies:[3]

(5)	A: We're all mad, aren't we?	$\forall x : M(x)$
	B: Well, some of us.	$\exists x : M(x)$
	$\rightsquigarrow$ *not (necessarily) all of us*	$\rightsquigarrow \exists x : \neg M(x)$
(6)	A: Check to see if your steak's burning.	$B(s)$
	B: Well something's bloody burning.	$\exists x : B(x)$
	$\rightsquigarrow$ *not (necessarily) my steak*	$\rightsquigarrow \neg B(s)$
(7)	A: Maybe three days.	$t = 3$
	B: Three or four days.	$t = 3 \vee t = 4$
	$\rightsquigarrow$ *not (necessarily) three*	$\rightsquigarrow \neg(t = 3)$
(8)	A: [Abbreviations are used] now in narrative and dialogue.	$N \wedge D$
	B: Well, in dialogue it's fine.	D
	$\rightsquigarrow$ *not (necessarily) in narrative*	$\rightsquigarrow \neg N$

What is remarkable about these rejections is that they are not only consistent with their antecedent, but are in fact *informationally redundant*—they are mere implications of the antecedent and as such intuitively innocuous. On the other hand, it is unexpected that a contradicting implicature[4] can arise at all: Since implicatures can be cancelled by prior context, the occurrence of an inconsistent implicature is unexpected from a theoretical standpoint (Walker, 1996).

Already Horn (1989) observed that some rejections are not semantic in nature, leading him to coin the term *metalinguistic negation*. Examples include rejections of implicatures, as in (9) and (10), or of presuppositions as in (11):[5]

(9) A: It's your job.
$\rightsquigarrow$ *your job alone*
B: It's our job.

(10) A: Three or four days.
$\rightsquigarrow$ *exact value unknown*
B: Well, four.

(11) A: Put a special colour of the buttons.
$\rightsquigarrow$ *there are buttons*
B: But we don't have any buttons.

[3] Examples from the BNC (Burnard, 2000).

[4] Walker (1996) called these *implicature rejection*; we cannot adopt the terminology, as we need to discern rejection-by-implicature from rejection-of-implicature below.

[5] Examples (9) and (11) from the AMI Corpus (Carletta, 2007), and (10) from the BNC (Burnard, 2000).

Rejections of implicatures are also (semantically) consistent with their antecedent, though they need not be informationally redundant, and rejections of presuppositions only become inconsistent once the presupposition has been computed. These examples are also not accounted for by the standard approach: neither can their rejection force be determined by a simple search for inconsistency, nor do these rejections amount to asserting the negative of their antecedent. In (9), B cannot be said to assert that it is not her job, she just takes offence to the connotation that it is her's alone, and in (11), B also cannot be taken to assert that there should *not* be a special colour on the buttons. An interesting case of rejections-of-implicatures are utterances that are taken to be more general than their addressee is willing to accept. In the following examples, the offending implicature arises because B expected (or wanted) A to be more specific; the lack of specificity gives rise to a generalising implicature:[6]

(12) A: You like country. $\rightsquigarrow$ *country in general*
B: But not all country.

(13) A: You love soap. $\rightsquigarrow$ *soaps in general*
B: I love lovely soaps. $\rightsquigarrow$ *not all soaps*

Example (13) is a particular case where it is both an implicature that is rejected, and an implicature that does the rejecting: A's utterance is pragmatically enriched to a general interpretation by B, exactly as in (12), but instead of explicitly rejecting this, A *implicates* the rejecting *'not all'*.

In general, when we speak of a *rejection* (or more generally of *rejection force*), we mean an answer to an earlier assertion or proposal that signals the speaker's refusal to add the assertion/proposal's content to the common ground. In particular, we exclude replies to questions from our study, since a negative answer to a polar question has a different influence on the common ground: it adds the negative counterpart of the question's propositional content. From now on we say that an utterance is a *pragmatic rejection* if it has rejection force, but is semantically consistent with what it is rejecting. We restrict our discussion to the three types exemplified above: rejection-by-implicature, rejection-of-implicature and rejection-of-presupposition.[7] We are concerned with the task of determining which utterances are (pragmatic) rejections, *i.e.*, given a (consistent) utterance, how can we determine if it has rejecting force?

3 Related Work

A typical area of interest for rejection computing is the summarisation of multiparty meeting transcripts (Hillard et al., 2003; Hahn et al., 2006; Schlöder and Fernández, 2014) and online discussions (Yin et al., 2012; Misra and Walker, 2013). This body of work has identified a number of local and contextual features that are helpful in discerning agreement from disagreement. However, their models—if explicated—rarely take pragmatic functions into account. Also, the task of retrieving rejections remains a computational challenge; Germesin and Wilson (2009) report high accuracy in the task of classifying utterances into *agreement / disagreement / other*, but have 0% recall of disagreements.

Walker (1996, 2012) first raised the issue of *implicature rejection*. Building on Horn's (1989) landmark exposition on negation and his discussion of *metalinguistic* rejections, she describes a new class of utterances which are not in contradiction with their antecedent, but nevertheless have rejection force. Her prototypical example is *'A: There's a man in the garage.' – 'B: There's something in the garage.'* (similar to our (6)), where B's utterance is clearly implied by A's, but rejecting by virtue of a Quantity implicature. An example where the rejecting speaker adds a disjunction, similar to our (7) (*'A: Maybe three days.' – 'B: Three or four days.'*), was already discussed by Grice (1991, p. 82), though he did not make an explicit connection to rejections *by* implicatures,[8] even though he mentions that there are rejections *of* implicatures. Walker (2012) is concerned with the question of why a rejecting implicature is not cancelled by prior context, and proposes to stratify the grounding process into different levels of *endorsement*, where a tacit endorsement does not have cancellation force.

[6] Example (12) from the Switchboard Corpus (Godfrey et al., 1992) and (13) from the BNC (Burnard, 2000).

[7] We do not claim that this is an exhaustive categorisation. In particular, we think that rejection-by-presupposition is also possible, as in the constructed example *'A: John never smoked.' – 'B: He stopped smoking before you met him.'*

[8] His (mostly unrelated) discussion centres on the semantics and pragmatics of the conditional *'if.'* He remarks in passing that replying *'X or Y or Z'* to *'X or Y'* rejects the latter, although "not as false but as *unassertable*" (his emphasis).

Potts (2011) and de Marneffe et al. (2009) have investigated a phenomenon similar to pragmatic rejection: They study answers to polar questions which are *indirect* in that they do not contain a clear *'yes'* or *'no'* and therefore their intended polarity must be inferred—sometimes by pragmatic means. They describe answers that require linguistic knowledge—such as salient scales—to be resolved; these are similar to our examples (3) and (4). Potts (2011) reports the results of a crowdsourcing experiment where participants had to judge whether an indirect response stood for a *'yes'* or a *'no'* answer. He then analyses indirect responses by their relative *strength* compared to the question radical. His experimental data shows that a weaker item in the response generally indicates a negative answer (*'A: Did you manage to read that section I gave you?' – B: 'I read the first couple of pages.'*), while a stronger item in the response generally indicates a positive answer (*'A: Do you like that one?' – 'B: I love it.'*). The former result corresponds to our rejection-by-implicature, while the latter is in contrast to our intuitions on rejection-of-implicature. As mentioned, our focus lies with rejections of assertions rather than answers to polar questions. Since the results of Potts and colleagues do not straightforwardly generalise from polar questions to assertions, we have adapted their methodology to conduct a study on responses to assertions; we return to this in Section 6.

4 A Corpus of Pragmatic Rejections

To our knowledge, there is currently no corpus available which is suitable to investigate the phenomenon of pragmatic rejection. We assembled such a corpus from three different sources: the AMI Meeting Corpus (Carletta, 2007), the Switchboard corpus (Godfrey et al., 1992) and the spoken dialogue section of the British National Corpus (Burnard, 2000). Since, generally, rejection is a comparatively rare phenomenon,[9] pragmatic rejections are few and far between. As indicated above, we consider an utterance a *pragmatic rejection* if it has rejection force, but is not (semantically) in contradiction to the proposal it is rejecting. As it is beyond the current state of the art to computationally search for this criterion, our search involved a substantial amount of manual selection. We assembled our corpus as follows:

- The AMI Meeting Corpus is annotated with relations between utterances, loosely called *adjacency pair* annotation.[10] The categories for these relations include `Objection/Negative Assessment (NEG)` and `Partial Agreement/Support (PART)`. We searched for all `NEG` and `PART` adjacency pairs where the first-part was *not* annotated as a question-type (`Elicit-*`) dialogue act, and manually extracted pragmatic rejections.
- The Switchboard corpus is annotated with dialogue acts,[11] including the tags `ar` and `arp` indicating (partial) rejection. We searched for all turn-initial utterances that are annotated as `ar` or `arp` and manually extracted pragmatic rejections.
- In the BNC we used SCoRE (Purver, 2001) to search for words or phrases repeated in two adjacent utterances, where the second utterance contains a rejection marker like *'no'*, *'not'* or turn-initial *'well'*; for repetitions with *'and'* in the proposal or *'or'* in the answer; for repetitions with an existential quantifier *'some*'* in the answer; for utterance-initial *'or'*; and for the occurrence of scalar implicatures by manually selecting scales and searching for the adjacent occurrence of different phrases from the same scale, *e.g.*, *'some – all'* or *'cold – chilly'*. We manually selected pragmatic rejections from the results.

Using this methodology, we collected a total of 59 pragmatic rejections. We categorised 16 of those as rejections-of-implicature, 33 as rejections-by-implicature, 4 as both rejecting *an* implicature and rejecting *by* one, and 6 as rejections-of-presupposition. All examples used in Section 2 are taken from our

[9] Schlöder and Fernández (2014) report 145 rejections of assertions in the Switchboard, and 679 in the AMI; as the BNC contains mainly free conversation, rejections are expected to be rare dispreferred acts (Pomerantz, 1984). We also note that Walker (1996) did not report any 'implicature rejections' or rejections-of-presupposition from her dataset.

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

[11] See `http://web.stanford.edu/~jurafsky/ws97/manual.august1.html`.

corpus.[12] While this small corpus is the first collection of pragmatic rejections we are aware of, we note that it is ill-suited for quantitative analysis: On one hand, we cannot be sure that the annotators of the Switchboard and AMI corpora were aware of pragmatic rejection and therefore might have not treated them uniformly—in fact, that we found pragmatic rejections in the AMI annotated as `NEG` and as `PART` supports this. On the other hand, our manual selection might not be balanced, since we cannot make any claims to have surveyed the corpora, particularly the BNC, exhaustively. In particular, that we did not find any rejections-by-presupposition should not be taken as an indication that the phenomenon does not occur. While we did find some rejections related to *conventional* implicatures,[13] we did not include them in our corpus as they are not the subject of our study.

5 Computing Rejections-by-Implicature

In this section we focus on the rejections-*by*-implicature. We contend that rejections-of-implicatures and rejections-of-presuppositions can be adequately captured by standard means to compute implicatures and presuppositions. For example in van der Sandt's (1994) model, a rejection can address the whole informational content, including pragmatic enrichment, of its antecedent utterance. However, it is a challenge to formally determine the rejection force of a rejection-by-implicature (Walker, 2012). We present a formal model that accounts for rejections-by-implicature and directly generalises on approaches that stipulate *rejection as inconsistency*. As a proof-of-concept, we discuss an implementation of our model in a probabilistic framework for simulation.

5.1 Formal Detection

The examples for rejection-by-implicature we found share some common characteristics: they are all informationally redundant, and they are all rejections by virtue of a Quantity implicature (Grice, 1975).[14] The crucial observation is that they are in fact not only informationally redundant, but are *strictly less* informative than their antecedent utterance, if we were to consider both individually. Recall the examples from Section 2, *e.g.*:

(8)	A: [Abbreviations are used] now in narrative and dialogue.	$N \wedge D$
	B: Well, in dialogue it's fine.	D
	$\rightsquigarrow$ *not (necessarily) in narrative*	$\rightsquigarrow \neg N$

Now, the Quantity implicature can be explained by the loss of information: The less informative utterance expresses that the speaker is unwilling to commit to any stronger statement. However, this still leaves open why the implicature is not cancelled by the prior context. The answer is deceptively simple: The rejecting speaker does not ground the prior utterance—does not make it part of her individual representation of context—and hence it has no influence on the implicatures in her utterance. This is deceptive because this explanation puts us in a chicken-and-egg situation: The implicature can arise only because the utterance is rejecting, but the utterance is rejecting because of the implicature.

The key to the solution is the informational redundancy. An informationally redundant utterance serves *a priori* no discourse function, therefore some additional reasoning is required to uncover the speaker's meaning (Stalnaker, 1978). In particular, an utterance may *appear* to be informationally redundant, but only because the speaker's context has been misconstrued: If we attribute too narrow a context to the utterance, it might just *seem* redundant. Hence we propose the following: If an utterance is informationally redundant, its informational content should be evaluated in the (usually wider) *prior context*, *i.e.*, in the context where the preceding utterance was made. If, then, the utterance turns out to be less informative than its antecedent, it is a pragmatic rejection. We call the enlargement of the context

[12] The corpus, our classification, as well as the results of our experiment described in Section 6, will be made freely available.

[13] From the BNC: *'They are not sandals, they're flip-flops'* when discussing the appropriateness of someone's attire.

[14] In principle, rejections by Quality or Relation implicatures seem possible: A Quality implicature could arise if someone says something absurd, which our model would consider a rejection by semantic means. A sudden change of topic, flouting the Relation Maxim, might be used as a rejection. However, detecting topic changes is far beyond the scope of our work.

the *pragmatic step*. Note that the pragmatic step itself makes no reference to any implicatures or to the rejection function or the utterance, thereby avoiding the chicken-and-egg problem.

We claim that this easily extends current models that adhere to a *rejection as inconsistency* paradigm. As a demonstration, we present the appropriate account in possible world semantics. Let $[\![\cdot]\!]_\cdot$ stand for the context update function, mapping sets of possible worlds to smaller such sets: $[\![u]\!]_c$ is the information state obtained by uttering u in c. We now describe when utterance u_2 rejects the previous utterance u_1. For brevity, write c_1 for the context in which u_1 is uttered, and $c_2 = [\![u_1]\!]_{c_1}$ for u_2's context. Then we can attempt a definition:

$$u_2 \text{ rejects } u_1 \text{ iff } ([\![u_2]\!]_{c_2} = \emptyset) \vee ([\![u_2]\!]_{c_2} = c_2 \wedge [\![u_2]\!]_{c_1} \supsetneq [\![u_1]\!]_{c_1}).$$

That is, u_2 has rejecting force if it is a plain inconsistency (reducing the context to absurdity), or if it is informationally redundant (does not change the context) and is properly less informative than its antecedent (would result in a larger context set if uttered in the same place). If we stipulate that the context update function captures pragmatic enrichment, *i.e.*, computes implicatures and presuppositions, then we capture the other pragmatic rejections by the inconsistency condition.

However, a technicality separates us from the complete solution: The rejecting utterance u_2 might be—and frequently is—non-sentential and/or contain pronominal phrases relating to u_1. That means that it actually cannot be properly interpreted in the prior context: the informational content of u_1 is required after all. Consider for example the following rejection-by-implicature:

(14)	A: Four. Yeah.	$x = 4$
	B: Or three.	$x = 4 \vee x = 3$
	$\rightsquigarrow$ *not (necessarily) four*	$\rightsquigarrow \neg(x = 4)$

Here, B's utterance requires the contextual information of A's previous turn to have the meaning *'four or three.'* To account for this, we need to separate the context into a *context of interpretation* (the discourse context, including everything that has been said) and a *context of evaluation* (the information against which the new proposition is evaluated) and only do the pragmatic step on the evaluative context. Now, the full paradigm for rejection in possible world semantics reads as:

$$u_2 \text{ rejects } u_1 \text{ iff } ([\![u_2]\!]_{d_2,e_2} = \emptyset) \vee ([\![u_2]\!]_{d_2,e_2} = e_2 \wedge [\![u_2]\!]_{d_2,e_1} \supsetneq [\![u_1]\!]_{d_1,e_1}).$$

Where d_1 and d_2 are the interpretative contexts in which u_1 and u_2 are uttered, respectively, and e_1 and e_2 are the corresponding evaluative contexts. Here, $[\![u]\!]_{d,e}$ maps an utterance u, an interpretative context d and an evaluative context e to an evaluative context: The context obtained by interpreting u in d and updating e with the result.

This is not a new—or particularly surprising—approach to context. Already Stalnaker (1978) proposed a two-dimensional context to discern interpretation from evaluation, though his concern was not mainly with non-sentential utterances, but rather with names and indexicals. However, more recent theories of dialogue semantics employing *structured contexts*, *e.g.*, KoS (Ginzburg, 2012) or PTT (Poesio and Traum, 1997), make frequent use of multi-dimensional representations of context to solve problems of anaphora resolution or the interpretation of non-sentential utterances. Typically, such systems keep track of what is *under discussion* separate from the *joint beliefs* and use the former to facilitate utterance interpretation. This roughly corresponds to our separation of interpretative and evaluative context.

This characterisation of rejection describes all semantic rejections, understood as inconsistencies, and adds the rejections-by-implicature via the pragmatic step. It does not overcommit either: An acceptance, commonly understood, is either more informative than its antecedent (a relevant next contribution), or informationally redundant when mirroring the antecedent utterance,[15] but then not *less* informative in the prior context. This includes the informationally redundant acceptance which puzzled Walker (1996):

(15) A: Sue's house is on Chestnut St.
B: on Chestnut St.

[15] Either by repeating a fragment of the antecedent, or by a particle like *'yes'*, which is understood to pick up the antecedent.

Walker (1996) claims that (15) is informationally redundant and less informative than the antecedent, hence it is expected to be an implicature rejection—but factually is a confirmation. Our model solves the issue: If B's non-sentential utterance is enriched by the interpretative context in the aftermath of A's utterance, it has *exactly* the informational content of its antecedent, and therefore is correctly predicted to be accepting.

5.2 Computational Simulation

The probabilistic programming language Church (Goodman et al., 2008) has been put forward as a suitable framework to model pragmatic reasoning. We have implemented our formal model on top of an implementation of Quantity implicatures by Stuhlmüller (2014). The implementation models two classically Gricean interlocutors: Speakers reason about rational listener behaviour and vice versa. Stuhlmüller's (2014) original model simulated scalar implicatures; we adapted his model to capture the *'and'/'or'* Quantity implicatures of examples (7) and (8).

The world in our model has two states, p and q, that can each be true or false. The speaker's vocabulary is {`neither, p, q, not-p, not-q, p-or-q, p-and-q`}. The listener guesses the state of the world as follows: Given a message, the listener reasons by guessing a rational speaker's behaviour given a rational listener; 'guessing' is done via sampling functions built into the programming language. For example, the message `p-and-q` unambiguously communicates that $p \wedge q$, because no rational listener would conclude from `p-and-q` that $\neg p$ or $\neg q$. On the other hand, `p-or-q` is taken to communicate $p \wedge \neg q$ and $\neg p \wedge q$ in about 40% of samples each and $p \wedge q$ in about 20% of samples, because all three states are consistent with the message, but a rational speaker would rather choose `p-and-q` in $p \wedge q$ because it is unambiguous.

The message `p` induces the belief that $p \wedge \neg q$ in roughly 65% of samples, and $p \wedge q$ in the remaining 35%. Again this is due to the fact that $p \wedge \neg q$ is indeed best communicated by `p`, whereas $p \wedge q$ is better communicated by `p-and-q`—the listener cannot exclude that q holds but thinks it less likely than $\neg q$ because different speaker behaviour would be expected if q were true.

Message	Response	Rejection
`p`	`not-p`	100%
`p-and-q`	`not-p`	100%
`p`	`p`	0%
`p-or-q`	`p-and-q`	0%
`p-or-q`	`p`	0%
`p`	`p-and-q`	0%
`p-and-q`	`p-or-q`	**78%**
`p-and-q`	`p`	**65%**
`p`	`p-or-q`	**64%**
`p`	`q`	59%

Table 1: Probabilities (1000 samples) that a pragmatically reasoning speaker would recognize a rejection according to our model.

For the implementation of rejection-by-implicature, we proceed as follows: Given a proposal and a response, obtain a belief by sampling a state of the world consistent with rational listener behaviour when interpreting the *response*: this is evaluating the response in the prior context, *i.e.*, computing what would happen if the speaker would utter the response instead of the proposal. Then check if this belief could have also been communicated by the proposal; if *not*, then the response is less informative (because it is consistent with more beliefs) than the proposal, and the model judges the response as rejecting. In each sample, the model makes a binary choice on whether it judges the response as rejecting or accepting.

We report some test runs of the simulation in Table 1, where for each proposal–response pair we computed 1000 samples. We observe that semantic rejections (*i.e.*, inconsistencies) are assigned rejection force with 100% confidence, and utterances intuitively constituting acceptances are never considered rejections. Some of the acceptances might be rejections-of-implicatures (being strictly more informative than the message they are replying to), but since our model does not pragmatically enrich the message, these are not found. In fact, it is not clear to us when, without further context or markers, replying `p` to `p-or-q` is an acceptance-by-elaboration or a rejection-of-implicature.[16] For now, the model considers this a clear acceptance: The implicature `p-or-q` $\rightsquigarrow \neg p$ needs to be computed elsewhere if at all.

Implicature rejections are not assigned 100% rejection force due to the probabilistic model for pragmatic reasoning. Since, per the model, the utterance `p-or-q` induces the belief that $p \wedge q$ in at least some

[16] We hypothesise that more subtle cues are required to make the distinction; this also fits our experimental results below.

In the following dialogues, speaker A makes a statement and speaker B reacts to it, but rather than simply agreeing or disagreeing by saying Yes/No, B responds with something more indirect and complicated. For instance:

A: It looks like a rabbit.
B: I think it's like a cat.

Please indicate which of the following options best captures what speaker B meant in each case:

- B definitely meant to agree with A's statement.
- B probably meant to agree with A's statement.
- B definitely meant to disagree with A's statement.
- B probably meant to disagree with A's statement.

In the sample dialogue above the right answer would be "B definitely meant to disagree with A's statement."

Cautionary note: in general, there is no unique right answer. However, a few of our dialogues do have obvious right answers, which we have inserted to help ensure that we approve only careful work.

Figure 1: CrowdFlower prompt with instructions, adapted from Potts (2011).

samples, the listeners cannot always recognize that `p-or-q` is an implicature rejection of `p-and-q`. In fact, the confidence that something is an implicature rejection scales with how well—how often—the implicature itself is recognized. Replying `q` to `p` is computed to be a rejection, because in the model `q` implicates that $\neg p$, as every utterance is taken to be about the state of both p and q.[17] In fact, replying `q` to `p` is also a rejection-*of*-implicature, as `p` also implicates $\neg q$. However, as mentioned above, our model does not capture this.

6 Annotation Experiment

In order to investigate the perceived rejection force of pragmatic rejections, we conducted an online annotation experiment using the corpus described in Section 4.

6.1 Setup

We adapted and closely followed the experimental setup of Potts (2011). The annotators were asked to rank the dialogues in our corpus on a 4-point scale: *'definitely agree'*, *'probably agree'*, *'probably disagree'* and *'definitely disagree'*. The instructions given to the participants, recruited on the crowdsourcing platform CrowdFlower,[18] are recorded in Figure 1. Like Potts (2011), we curated our corpus for the purposes of the annotation experiment by removing disfluencies and agrammaticalities to ensure that the participants were not distracted by parsing problems, as well as any polarity particles (including *'yeah'* and *'no'*) in the response utterance.

To ensure the quality of the annotation, we included some agreements and semantic disagreements as control items in the task.[19] Participants who ranked a control agreement as disagreeing or vice versa were excluded from the study. Some control items were chosen to require a certain amount of competence in discerning agreement from disagreement. For example, (16) is an agreement despite the negative polarity particle *'no'* appearing, and (17) requires an inference step with some substantial linguistic knowledge; (18) and (19) are examples for clear-cut agreement and disagreement respectively.

(16) A: I think wood is not an option either.
B: No, wood's not an option.

(17) A: We can't fail.
B: We fitted all the criterias.

(18) A: It's a giraffe.
B: A giraffe okay.

(19) A: Yes, a one.
B: I say a two.

[17] Due to this closed world assumption, we cannot say that this is a Relevance implicature.

[18] `http://www.crowdflower.com`

[19] Drawn from the AMI Corpus from items annotated as `Positive Assessment` and `Negative Assessment` respectively.

Rejection-	by-implicature					of-implicature			both	of-presupp.	**Total**
	or	*and*	*generalise*	*restrict*	*scalar*	*or*	*generalise*	*scalar*			
Raw number	12	5	2	3	11	1	8	7	4	6	59
Judged disagreeing	58%	17%	0%	26%	51%	21%	61%	42%	40%	68%	47%
Std. deviation	0.31	0.22	0	0.10	0.38	–	0.30	0.35	0.34	0.37	0.34

Table 2: Average percentage of *'probably/definitely disagreeing'* judgements by category.

We added 20 control agreements and 10 control disagreements to our corpus of pragmatic rejections, and presented each participant 9 dialogues at a time: 6 pragmatic rejections and 3 control items. Thereby we constructed 10 sets of dialogues, each of which was presented to 30 different participants. We filtered out participants who failed any of our control items from the results. The amount of filtered judgements was as high as 33% on some items. Polarity reversals like (16) were particularly effective in filtering out careless participants: Failure to recognise a polarity reversal shows a lack of contextual awareness, which is vital to judge pragmatic rejections.

6.2 Results and Discussion

For each item, we computed the percentage of participants who judged it as having rejecting force, *i.e.*, as either *'probably disagree'* or *'definitely disagree'*; see Table 2 for an overview of the results by category. To better understand our results, we classified the rejections-by/of-implicatures further by the implicature that gives rise to the rejection. We found the following sub-types in our dataset:

Rejections by means of an implicature:

- *or*-implicature as in (7): *'A: Maybe three days.' – 'B: Three or four days.'*
- *and*-implicature as in (8): *'A: [. . .] in narrative and dialogue.' – 'B: Well, in dialogue.'*
- *generalising* implicature as in (6): *'A: . . . your steak's burning.' – 'B: Well, something's burning.'*
- *restricting* implicature as in (5): *'A: We're all mad.' – 'B: Some of us.'*[20]
- *scalar* implicature as in (3): *'A: That's brilliant.' – 'B: [it] was quite good.'*

Rejections of an implicature:

- *or*-implicature as in (10): *'A: Three or four days.' – 'B: Well, four.'*
- *generalising* implicature as in (12): *'A: You like country' – 'B: But not all country.'*
- *scalar* implicature as in (4): *'A: It was good weren't it?' – 'B: It's brilliant.'*

Overall, about half of all judgements we collected deemed an item to have rejection force. These judgements were again split roughly 50-50 into *'probably disagree'* and *'definitely disagree.'* When a judgement did not indicate rejection force, *'probably agree'* was the preferred category, chosen in 78% of *'agree'* judgements. However, we saw substantial variation in the judgements when categorising the pragmatic rejections as above.[21]

Most notably, the two rejections by generalising implicature were never judged to have rejection force. Our hypothesis is that this is due to the fact that the surface form of these implicatures repeats some central phrase from their antecedent, and they are therefore taken to agree *partially*, which leads them to be judged as *'probably agree.'* For example, in the rejection by a generalising implicature (6), the interlocutors are apparently considered to agree on 'something *burning*.' The same observation holds for rejections by *and*-implicature, *e.g.*, in (8) the interlocutors might be judged to agree on the 'usage *in dialogue*.' In contrast, rejections by *or*-implicature and by scalar implicature stand out as being judged disagreeing more often: 58% and 51%, respectively. In our corpus, the surface form of such implicatures does not typically involve the repetition of a phrase from their antecedent. As a case in point, the rejection

[20] While this example could technically be considered a scalar implicature, we take *all-some* to be a special case of removing information; one can also restrict by adding adjectives to disagree with a universal statement, as in (13): *'A: You love [all] soap.' – 'B: I love lovely soaps.'*

[21] In contrast, we could not find any relation between our experimental results and previous annotations of the utterances in our corpus (if they were previously annotated, *i.e.*, taken from the AMI or Switchboard corpora).

by *or*-implicature (14) *'A: Four. Yeah.' – 'B: Or three.'* was judged to have rejection force much more frequently (86%) than the similar (7) *'A: Maybe three days.' – 'B: Three or four days.'* (40%) where B repeats part of A's proposal.[22] We think that other linguistic cues from the utterances' surface forms also had an influence on the perceived force of the responses. In particular, we attribute the high percentage of judged disagreements in the rejections of generalising implicatures (61%) to them being typically marked with the contrast particle *'but'*—a well known cue for disagreement (Galley et al., 2004; Misra and Walker, 2013; Schlöder and Fernández, 2014).

The rejections-of-presuppositions received the overall largest amount of rejection force judgements (68%). This is in accordance with previous work that has treated them in largely the same way as typical explicit rejections (Horn, 1989; van der Sandt, 1994; Walker, 1996). In particular, all rejections-of-presuppositions in our corpus correspond to utterances annotated as `Negative Assessment` in the AMI Corpus. That even these utterances received a substantial amount of *'probably agree'* judgements puts the overall results into context: The subjects show a noticeable tendency to choose this category.

The experimental results in Table 2 should not be compared quantitatively with the simulation outcome in Table 1, Section 5.2. The judgement scale in the experiment is in no direct relation with the probabilistic reasoning in the simulation. Qualitatively speaking, however, the experiment shows a difference in how rejections by *or*- and *and*-implicatures are perceived, whereas the simulation yields nigh-identical results for these two. This could be due to linguistic cues simply not present in the simulation, and due to participants in the experiment choosing *'probably agree'* when they perceived *partial* agreement in a dialogue. In contrast to such 'partial' judgements, our formal model considers agreement/disagreement as a binary distinction and infers full disagreement from slight divergences in informational content. We conclude from the experiment that this binary assumption should be given up, also in the probabilistic implementation, where the probabilities represent uncertainty about the world rather than the kind of partial agreement/disagreement that seems to be behind our experimental results.

7 Conclusion and Further Work

We have laid out the phenomenon of pragmatic rejection, given it a general definition, and assembled a small corpus of such rejections. Our formal model improves over extant work by capturing rejections-by-implicature. A simulation of the model has shown that it yields theoretically desirable results for agreements and semantic disagreements and predicts rejection force of rejections-by-implicature. Compared to our annotation experiment, however, the model lacks sophistication in computing what is apparently perceived as partial agreement/disagreement. The pragmatic rejections we collected were judged to have rejection force only about half of the time, and otherwise our subjects showed a preference for the category *'probably agree.'* We tentatively attribute this to linguistic cues, related to the surface form of some pragmatic rejections, which led the annotators to consider them partial agreements. We leave a deeper investigation into these cues to further work

In sum, while our model accounts for more data than previous approaches, we conclude that a more sophisticated model for rejection should give up the agree/disagree binary and account for utterances that fall inbetween; the data and analysis we presented here should be helpful to guide the development of such a model. Computing partial rejection force, particularly *which part* of an antecedent has been accepted or rejected, is part of our ongoing work.

References

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.

[22] The hedging *'maybe'* in A's utterance might also had an influence.

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

Galley, M., K. McKeown, J. Hirschberg, and E. Shriberg (2004). Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of ACL'04*.

Germesin, S. and T. Wilson (2009). Agreement detection in multiparty conversation. In *Proceedings of the 2009 international conference on Multimodal interfaces*.

Ginzburg, J. (2012). *The Interactive Stance*. Oxford University Press.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP'92*.

Goodman, N. D., V. K. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum (2008). Church: a language for generative models. In *Proceedings of Uncertainty in Artificial Intelligence*.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics,* Vol. 3, pp. 41–58. Acad. Press.

Grice, H. P. (1991). *Studies in the Way of Words*. Harvard University Press.

Hahn, S., R. Ladner, and M. Ostendorf (2006). Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of HLT-NAACL 2006*.

Hillard, D., M. Ostendorf, and E. Shriberg (2003). Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of HLT-NAACL 2003*.

Horn, L. R. (1989). *A Natural History of Negation*. University of Chicago Press.

Lascarides, A. and N. Asher (2009). Agreement, disputes and commitments in dialogue. *Journal of Semantics 26*(2), 109–158.

de Marneffe, M.-C., S. Grimm, and C. Potts (2009). Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference*.

Misra, A. and M. Walker (2013). Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGdial 2013 Conference*.

Poesio, M. and D. Traum (1997). Conversational actions and discourse situations. *Computational Intelligence 13*(3), 309–347.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In *Structures of Social Action*. Cambridge University Press.

Potts, C. (2011). The indirect question-answer pair corpus. `http://compprag.christopherpotts.net/iqap.html`. Accessed: 2014-11-24.

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

van der Sandt, R. A. (1994). Denial and negation. *Unpublished manuscript, University of Nijmegen*.

Schlöder, J. J. and R. Fernández (2014). The role of polarity in inferring acceptance and rejection in dialogue. In *Proceedings of the SIGdial 2014 Conference*.

Stalnaker, R. (1978). Assertion. In *Syntax and Semantics,* Vol. 9, pp. 315–332. Academic Press.

Stuhlmüller, A. (2014). Scalar Implicature. `http://forestdb.org/models/scalar-implicature.html`. Accessed: 2014-11-24.

Walker, M. A. (1996). Inferring acceptance and rejection in dialogue by default rules of inference. *Language and Speech 39*(2-3), 265–304.

Walker, M. A. (2012). Rejection by implicature. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*.

Yin, J., P. Thomas, N. Narang, and C. Paris (2012). Unifying Local and Global Agreement and Disagreement Classification in Online Debates. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*.

Clarifying Intentions in Dialogue: A Corpus Study

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 dealing with *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 discussed in terms of 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. We end with pointers for future work.

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

	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 understand 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.

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. 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*:[1]

(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:[2]

(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.[3]

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 ground-

[1] 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.

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

[3] We will make our annotated corpus freely available.

ing at the level of intentions.[4] 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,[5] which indicates whether or not an utterance is considered a direct reply to another one. This allowed us 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.[6]

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 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

[4] 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.

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

[6] 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.

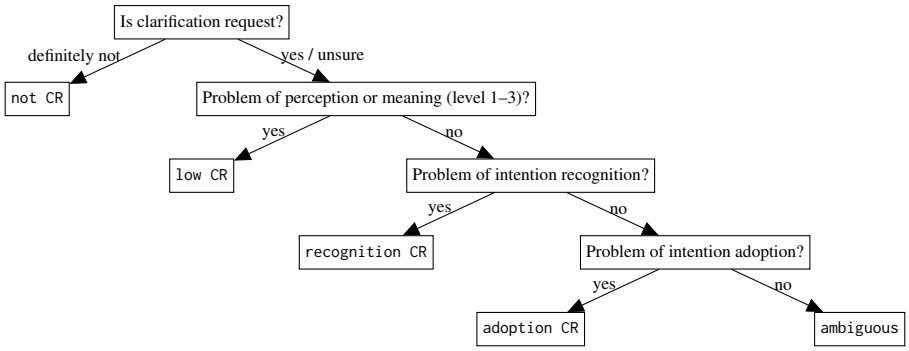

Figure 1: Decision Tree for Annotation of Clarification Requests.

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 the decision tree in Figure 1. 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. 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.

Since CRs can be fragmented and ambiguous, 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. This has also been observed by Rodríguez and Schlangen (2004) who include the CR asker's '`happiness`' (as evinced by the follow-up) in their annotation scheme.

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,

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.

the annotators discussed the utterances where their initial judgement differed and mutually agreed on the appropriate annotation.

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;[7] 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. Previous studies have indicated a higher number: Purver (2004) reports about 4% and Rodríguez and Schlangen (2004) about 5.8%. It is to be expected that our heuristic misses some CRs, *e.g.*, ones that do not receive an answer, and its coverage is also dependent on how systematic the adjacency pair annotation in the AMI Corpus is. 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). Despite this difference, our distribution of classes 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.[8] 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.

[7] 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.

[8] 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)*.

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

Buschmeier, H. and S. Kopp (2012). Using a bayesian model of the listener to unveil the dialogue information state. In *SemDial 2012: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*.

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)*.

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.