Implications for Generating Clarification Requests in Task-oriented Dialogues

Verena Rieser
Department of Computational Linguistics
Saarland University
Saarbrücken, D-66041
vrieser@coli.uni-sb.de

Johanna D. Moore
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, GB
J.Moore@ed.ac.uk

Abstract

Clarification requests (CRs) in conversation ensure and maintain mutual understanding and thus play a crucial role in robust dialogue interaction. In this paper, we describe a corpus study of CRs in task-oriented dialogue and compare our findings to those reported in two prior studies. We find that CR behavior in task-oriented dialogue differs significantly from that in everyday conversation in a number of ways. Moreover, the dialogue type, the modality and the channel quality all influence the decision of when to clarify and at which level of the grounding process. Finally we identify form-function correlations which can inform the generation of CRs.

1 Introduction

Clarification requests in conversation ensure and maintain mutual understanding and thus play a significant role in robust and efficient dialogue interaction. From a theoretical perspective, the *model of grounding* explains how mutual understanding is established. According to Clark (1996), speakers and listeners ground mutual understanding on four levels of coordination in an action ladder, as shown in Table 1.

Several current research dialogue systems can detect errors on different levels of grounding (Paek and Horvitz, 2000; Larsson, 2002; Purver, 2004; Schlangen, 2004). However, only the work of Purver (2004) addresses the question of how the source of the error affects the form the CR takes.

Level	Speaker S	Listener L
Convers.	S is proposing activity α	L is considering proposal α
Intention	S is signalling that *p*	L is recognizing that *p*
Signal	S is presenting signal σ	L is identifying signal σ
Channel	S is executing behavior β	L is attending to behavior β

Table 1: Four levels of grounding

In this paper, we investigate the use of form-function mappings derived from human-human dialogues to inform the generation of CRs. We identify the factors that determine which function a CR should take and identify function-form correlations that can be used to guide the automatic generation of CRs.

In Section 2, we discuss the classification schemes used in two recent corpus studies of CRs in human-human dialogue, and assess their applicability to the problem of generating CRs. Section 3 describes the results we obtained by applying the classification scheme of Rodriguez and Schlangen (2004) to the Communicator Corpus (Bennett and Rudnicky, 2002). Section 4 draws general conclusions for generating CRs by comparing our results to those of (Purver et al., 2003) and (Rodriguez and Schlangen, 2004). Section 5 describes the correlations between function and form features that are present in the corpus and their implications for generating CRs.

Proceedings of the 43rd Annual Meeting of the ACL, pages 239–246,
Ann Arbor, June 2005.

Attr.	Value	Category	Example
form	non	Non-Reprise	*"What did you say?"*
	wot	Conventional	*"Sorry?"*
	frg	Reprise Fragment	*"Edinburgh?"*
	lit	Literal Reprise	*"You want a flight to Edinburgh?"*
	slu	Reprise Sluice	*"Where?"*
	sub	Wh-substituted Reprise	*"You want a flight where?"*
	gap	Gap	*"You want a flight to...?"*
	fil	Gap Filler	*"...Edinburgh?"*
	other	Other	x
readings	cla	Clausal	*"Are you asking/asserting that X?"*
	con	Constituent	*"What do you mean by X?"*
	lex	Lexical	*"Did you utter X?"*
	corr	Correction	*"Did you intend to utter X instead?"*
	other	Other	x

Table 2: CR classification scheme by PGH

2 CR Classification Schemes

We now discuss two recently proposed classification schemes for CRs, and assess their usefulness for generating CRs in a spoken dialogue system (SDS).

2.1 Purver, Ginzburg and Healey (PGH)

Purver, Ginzburg and Healey (2003) investigated CRs in the British National Corpus (BNC) (Burnard, 2000). In their annotation scheme, a CR can take seven distinct surface forms and four readings, as shown in Table 2. The examples for the form feature are possible CRs following the statement *"I want a flight to Edinburgh"*. The focus of this classification scheme is to map semantic readings to syntactic surface forms. The *form* feature is defined by its relation to the problematic utterance, i.e., whether a CR reprises the antecedent utterance and to what extent. CRs may take the three different readings as defined by Ginzburg and Cooper (2001), as well as a fourth reading which indicates a correction.

Although PGH report good coverage of the scheme on their subcorpus of the BNC (99%), we found their classification scheme to to be too coarse-grained to prescribe the form that a CR should take. As shown in example 1, Reprise Fragments (RFs), which make up one third of the BNC, are ambiguous in their readings and may also take several surface forms.

(1) I would like to book a flight on Monday.

(a) Monday?
`frg, con/cla`

(b) Which Monday?
`frg, con`

(c) Monday the first?
`frg, con`

(d) The first of May?
`frg, con`

(e) Monday the first or Monday the eighth?
`frg, (exclusive) con`

RFs endorse literal repetitions of part of the problematic utterance (1.a); repetitions with an additional question word (1.b); repetition with further specification (1.c); reformulations (1.d); and alternative questions (1.e)[1].

In addition to being too general to describe such differences, the classification scheme also fails to describe similarities. As noted by (Rodriguez and Schlangen, 2004), PGH provide no feature to describe the extent to which an RF repeats the problematic utterance.

Finally, some phenomena cannot be described at all by the four readings. For example, the readings do not account for non-understanding on the pragmatic level. Furthermore the readings may have several problem sources: the clausal reading may be appropriate where the CR initiator failed to recognise the word acoustically as well as when he failed to resolve the reference. Since we are interested in generating CRs that indicate the source of the error, we need a classification scheme that represents such information.

2.2 Rodriguez and Schlangen (R&S)

Rodriguez and Schlangen (2004) devised a multi-dimensional classification scheme where *form* and

[1] Alternative questions would be interpreted as asking a polar question with an exclusive reading.

function are meta-features taking sub-features as attributes. The *function* feature breaks down into the sub-features *source, severity, extent, reply* and *satisfaction*. The *sources* that might have caused the problem map to the levels as defined by Clark (1996). These sources can also be of different *severity*. The severity can be interpreted as describing the set of possible referents: asking for repetition indicates that no interpretation is available (`cont-rep`); asking for confirmation means that the CR initiator has some kind of hypothesis (`cont-conf`). The *extent* of a problem describes whether the CR points out a problematic element in the problem utterance. The *reply* represents the answer the addressee gives to the CR. The *satisfaction* of the CR-initiator is indicated by whether he renews the request for clarification or not.

The meta-feature *form* describes how the CR is lingustically realised. It describes the *sentence's mood*, whether it is grammatically *complete*, the *relation to the antecedent*, and the *boundary tone*. According to R&S's classification scheme our illustrative example would be annotated as follows[2]:

(2) I would like to book a flight on Monday.

(a) Monday?

```
mood: decl
completeness: partial
rel-antecedent: repet
source: acous/np-ref
severity: cont-repet
extent: yes
```

(b) Which Monday?

```
mood: wh-question
completeness: partial
rel-antecedent: addition
source: np-ref
severity: cont-repet
extent: yes
```

(c) Monday the first?

```
mood: decl
completeness: partial
rel-antecedent: addition
source: np-ref
severity: cont-conf
extent: yes
```

(d) The first of May?

```
mood: decl
completeness: partial
rel-antecedent: reformul
source: np-ref
severity: cont-conf
extent: yes
```

(d) Monday the first or Monday the eighth?

```
mood: alt-q
completeness: partial
rel-antecedent: addition
source: np-ref
severity: cont-repet
extent: yes
```

In R&S's classification scheme, ambiguities about CRs having different sources cannot be resolved entirely as example (2.a) shows. However, in contrast to PGH, the overall approach is a different one: instead of explaining causes of CRs within a theoretic-semantic model (as the three different readings of Ginzburg and Cooper (2001) do), they infer the interpretation of the CR from the context. Ambiguities get resolved by the reply of the addressee and the satisfaction of the CR initiator indicates the "mutually agreed interpretation" .

R&S's multi-dimensional CR description allows the fine-grained distinctions needed to generate natural CRs to be made. For example, PGH's general category of RFs can be made more specific via the values for the feature *relation to antecedent*. In addition, the *form* feature is not restricted to syntax; it includes features such as intonation and coherence, which are useful for generating the surface form of CRs. Furthermore, the multi-dimensional *function* feature allows us to describe information relevant to generating CRs that is typically available in dialogue systems, such as the level of confidence in the hypothesis and the problem source.

3 CRs in the Communicator Corpus

3.1 Material and Method

Material: We annotated the human-human travel reservation dialogues available as part of the Carnegie Mellon Communicator Corpus (Bennett and Rudnicky, 2002) because we were interested in studying naturally occurring CRs in task-oriented dialogue. 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. The corpus comprises 31 dialogues of transcribed telephone speech, with 2098 dialogue turns and 19395 words.

[2]The source features answer and satisfaction are ignored as they depend on how the dialogue continues. The interpretation of the source is dependent on the reply to the CR. Therefore all possible interpretations are listed.

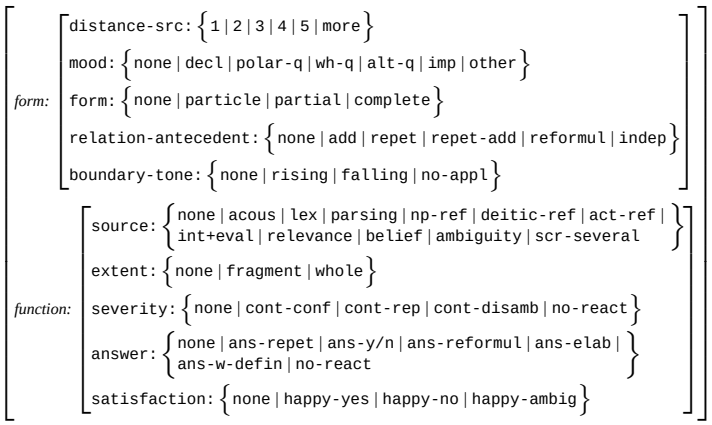

Figure 1: CR classification scheme

Annotation Scheme: Our annotation scheme, shown in Figure 1, is an extention of the R&S scheme described in the previous section. R&S's scheme was devised for and tested on the Bielefeld Corpus of German task-oriented dialogues about joint problem solving.[3] To annotate the Communicator Corpus we extended the scheme in the following ways. First, we found the need to distinguish CRs that consist only of newly added information, as in example 3, from those that add information while also repeating part of the utterance to be clarified, as in 4. We augmented the scheme to allow two distinct values for the *form* feature `relation-antecedent`, `add` for cases like 3 and `repet-add` for cases like 4.

(3) **Cust:** What is the last flight I could come back on?
Agent: On the 29th of March?

(4) **Cust:** I'll be returning on Thursday the fifth.
Agent: The fifth of February?

To the function feature *source* we added the values `belief` to cover CRs like 5 and `ambiguity refinement` to cover CRs like 6.

(5) **Agent:** You need a visa.
Cust: I *do* need one?
Agent: Yes you do.

(6) **Agent:** Okay I have two options . . . with Hertz . . . if not they do have a lower rate with Budget and that is fifty one dollars.
Cust: Per day?
Agent: Per day um mm.

Finally, following Gabsdil (2003) we introduced an additional value for *severity*, `cont-disamb`, to cover CRs that request disambiguation when more than one interpretation is available.

Method: We first identified turns containing CRs, and then annotated them with *form* and *function* features. It is not always possible to identify CRs from the utterance alone. Frequently, context (e.g., the reaction of the addressee) or intonation is required to distinguish a CR from other feedback strategies, such as positive feedback. See (Rieser, 2004) for a detailed discussion. The annotation was only performed once. The coding scheme is a slight variation of R&S, which has been shown relaiable with Kappa of 0.7 for identifying source.

3.2 Forms and Functions of CRs in the Communicator Corpus

The human-human dialogues in the Communicator Corpus contain 98 CRs in 2098 dialogue turns (4.6%).

Forms: The frequencies for the values of the individual *form* features are shown in Table 3. The most frequent type of CRs were *partial declarative questions*, which combine the mood value `declarative` and the completeness value `partial`.[4] These account for 53.1% of the CRs in the corpus. Moreover, four of the five most frequent surface forms of CRs in the Communicator Corpus differ only in the value for the feature `relation-antecedent`. They are partial declaratives with rising boundary tone, that either reformulate (7.1%) the problematic utterance, repeat

[3] `http://sfb360.uni-bielefeld.de`

[4] Declarative questions cover "all cases of non-interrogative word-order, i.e., both declarative sentences and fragments" (Rodriguez and Schlangen, 2004).

Feature	Value	Freq. (%)
Mood	declarative	65
	polar	21
	wh-question	7
	other	7
Completeness	partial	58
	complete	38
	other	4
Relation antecedent	rep-add	27
	independent	21
	reformulation	19
	repetition	18
	addition	10
	other	5
Boundary tone	rising	74
	falling	22
	other	4

Table 3: Distribution of values for the *form* features

Feature	Value	Freq. (%)
Source	np-reference	40
	acoustic	31
	intention	8
	belief	6
	ambiguity	4
	contact	4
	others	3
	relevance	2
	several	2
Extent	yes	80
	no	20
Severity	confirmation	73
	repetition	20
	other	7
Answer	y/n answer	64
	other	15
	elaboration	13
	no reaction	6

Table 4: Distribution of values for the *function* features

the problematic constituent (11.2%), add only new information (7.1%), or repeat the problematic constituent and add new information (10.2%). The fifth most frequent type is conventional CRs (10.2%).[5]

Functions: The distributions of the *function* features are given in Figure 4. The most frequent source of problems was np-reference. Next most frequent were acoustic problems, possibly due to the poor channel quality. Third were CRs that enquire about intention. As indicated by the feature *extent*, almost 80% of CRs point out a specific element of the problematic utterance. The features *severity* and *answer* illustrate that most of the time CRs request confirmation of an hypothesis (73.5%) with a yes-no-answer (64.3%). The majority of the provided answers were satisfying, which means that the addressee tends to interpret the CR correctly and answers collaboratively. Only 6.1% of CRs failed to elicit a response.

4 CRs in Task-oriented Dialogue

4.1 Comparison

In order to determine whether there are differences as regards CRs between task-oriented dialogues and everyday conversations, we compared our results to those of PGH's study on the BNC and those of R&S on the Bielefeld Corpus. The BNC contains a 10 million word sub-corpus of English dialogue transcriptions about topics of general interest. PGH analysed a portion consisting of ca. 10,600 turns, ca. 150,000 words. R&S annotated 22 dialogues from the Bielefeld Corpus, consisting of ca. 3962 turns, ca. 36,000 words.

The major differences in the feature distributions are listed in Table 5. We found that there are no significant differences between the feature distributions for the Communicator and Bielefeld corpora, but that the differences between Communicator and BNC, and Bielefeld and BNC are significant at the levels indicated in Table 5 using Pearson's χ^2. The differences between dialogues of different types suggest that there is a different grounding strategy. In task-oriented dialogues we see a trade-off between avoiding misunderstanding and keeping the conversation as efficient as possible. The hypothesis that grounding in task-oriented dialogues is more *cautious* is supported by the following facts (as shown by the figures in Table 5):

- CRs are more frequent in task-oriented dialogues.
- The overwhelming majority of CRs directly follow the problematic utterance.

[5] Conventional forms are *"Excuse me?"*, *"Pardon?"*, etc.

	Corpus		
Feature	Communicator	Bielefeld	BNC
CRs	98	230	418
frequency	4.6%	5.8%***	3.9%
distance-src=1	92.8%*	94.8%***	84.4%
no-react	6.1%*	8.7%**	17.0%
cont-conf	73.5%***	61.7%***	46.6%
partial	58.2%**	76.5%***	42.4%
independent	21.4%***	9.6%***	44.2%
cont-rep	19.8%***	14.8%***	39.5%
y/n-answer	64.3%	44.8%	n/a

Table 5: Comparison of CR *forms* in everyday vs. task-oriented corpora (* denotes $p < .05$, ** is $p < .01$, *** is $p < .005$.)

	Corpus		
Feature	Communicator	Bielefeld	Significance
contact	4.1%	0 inst	n/a
acoustic	30.6%	11.7%	***
lexical	1 inst	1 inst	n/a
parsing	1 inst	0 inst	n/a
np-ref	39.8%	24.4%	**
deict-ref	1 inst	27.4%	***
ambiguity	4.1%	not eval.	n/a
belief	6.1%	not eval.	n/a
relevance	2.1%	not eval.	n/a
intention	8.2%	22.2%	**
several	2.0%	14.3%	***

Table 6: Comparison of CR problem sources in task-oriented corpora

- CRs in everyday conversation fail to elicit a response nearly three times as often.[6]
- Even though dialogue participants seem to have strong hypotheses, they frequently confirm them.

Although grounding is more cautious in task-oriented dialogues, the dialogue participants try to keep the dialogue as *efficient* as possible:

- Most CRs are partial in form.
- Most of the CRs point out one specific element (with only a minority being independent as shown in Table 5). Therefore, in task-oriented dialogues, CRs locate the understanding problem directly and give partial credit for what was understood.
- In task-oriented dialogues, the CR-initiator asks to confirm an hypothesis about what he understood rather than asking the other dialogue participant to repeat her utterance.
- The addressee prefers to give a short y/n answer in most cases.

Comparing error sources in the two task-oriented corpora, we found a number of differences as shown in Table 6. In particular:

- *Dialogue type*: Belief and ambiguity refinement do not seem to be a source of problems in joint problem solving dialogues, as R&S did not include them in their annotation scheme. For CRs in information seeking these features need to be added to explain quite frequent phenomena. As shown in Table 6, 10.2% of CRs were in one of these two classes.
- *Modality*: Deictic reference resolution causes many more understanding difficulties in dialogues where people have a shared point of view than in telephone communication (Bielefeld: most frequent problem source; Communicator: one instance detected). Furthermore, in the Bielefeld Corpus, people tend to formulate more fragmentary sentences. In environments where people have a shared point of view, complete sentences can be avoided by using non-verbal communication channels. Finally, we see that establishing contact is more of a problem when speech is the only modality available.
- *Channel quality*: Acoustic problems are much more likely in the Communicator Corpus.

These results indicate that the decision process for grounding needs to consider the modality, the domain, and the communication channel. Similar extensions to the grounding model are suggested by (Traum, 1999).

[6] Another factor that might account for these differences is that the BNC contains multi-party conversations, and questions in multi-party conversations may be less likely to receive responses. Furthermore, due to the poor recording quality of the BNC, many utterances are marked as "not interpretable", which could also lower the response rate.

4.2 Consequences for Generation

The similarities and differences detected can be used to give recommendations for generating CRs. In terms of when to initiate a CR, we can state that clarification should not be postponed, and immediate, local management of uncertainty is critical. This view is also supported by observations of how non-native speakers handle non-understanding (Paek, 2003).

Furthermore, for task-oriented dialogues the system should present an hypothesis to be confirmed, rather than ask for repetition. Our data suggests that, when they are confronted with uncertainty, humans tend to build up hypotheses from the dialogue history and from their world knowledge. For example, when the customer specified a date without a month, the travel agent would propose the most reasonable hypothesis instead of asking a wh-question. It is interesting to note that Skantze (2003) found that users are more satisfied if the system "hides" its recognition problem by asking a task-related question to help to confirm the hypothesis, rather than explicitly indicating non-understanding.

5 Correlations between Function and Form: How to say it?

Once the dialogue system has decided on the *function* features, it must find a corresponding surface form to be generated. Many forms are indeed related to the function as shown in Table 7, where we present a significance analysis using Pearson's χ^2 (with Yates correction).

Source: We found that the relation to the antecedent seems to distinguish fairly reliably between CRs clarifying reference and those clarifying acoustic understanding. In the Communicator Corpus, for acoustic problems the CR-initiator tends to repeat the problematic part literally, while reference problems trigger a reformulation or a repetition with addition. For both problem sources, partial declarative questions are preferred. These findings are also supported by R&S. For the first level of non-understanding, the inability to establish contact, complete polar questions with no relation to the antecedent are formulated, e.g., "*Are you there?*".

Severity: The severity indicates how much was understood, i.e., whether the CR initiator asks to confirm an hypothesis or to repeat the antecedent utterance. The severity of an error strongly correlates with the sentence mood. Declarative and polar questions, which take up material from the problematic utterance, ask to confirm an hypothesis. Wh-questions, which are independent, reformulations or repetitions with additions (e.g., wh-substituted reprises) of the problematic utterance usually prompt for repetition, as do imperatives. Alternative questions prompt the addressee to disambiguate the hypothesis.

Answer: By definition, certain types of question prompt for certain answers. Therefore, the feature *answer* is closely linked to the sentence mood of the CR. As polar questions and declarative questions generally enquire about a proposition, i.e., an hypothesis or belief, they tend to receive yes/no answers, but repetitions are also possible. Wh-questions, alternative questions and imperatives tend to get answers providing additional information (i.e., reformulations and elaborations).

Extent: The *function* feature *extent* is logically independent from the *form* feature *completeness*, although they are strongly correlated. *Extent* is a binary feature indicating whether the CR points out a specific element or concerns the whole utterance. Most fragmentary declarative questions and fragmentary polar questions point out a specific element, especially when they are not independent but stand in some relation to the antecedent utterance. Independent complete imperatives address the whole previous utterance.

The correlations found in the Communicator Corpus are fairly consistent with those found in the Bielefeld Corpus, and thus we believe that the guidelines for generating CRs in task-oriented dialogues may be language independent, at least for German and English.

6 Summary and Future Work

In this paper we presented the results of a corpus study of naturally occurring CRs in task-oriented dialogue. Comparing our results to two other studies, one of a task-oriented corpus and one of a cor-

Form	Function			
	source	severity	extent	answer
mood	$\chi^2(24) = 112.20$ $p < 0.001$	$\chi^2(5) = 30.34$ $p < 0.001$	$\chi^2(5) = 24.25$ $df = p < 0.005$	$\chi^2(5) = 25.19$ $p < 0.001$
bound-tone	*indep.*	*indep.*	*indep.*	*indep.*
rel-antec	$\chi^2(24) = 108.23$ $p < 0.001$	$\chi^2(4) = 11.69$ $p < 0.005$	$\chi^2(4) = 42.58$ $p < 0.001$	*indep.*
complete	$\chi^2(7) = 27.39$ $p < 0.005$	*indep.*	$\chi^2(1) = 27.39$ $p < 0.001$	*indep.*

Table 7: Significance analysis for form/function correlations.

pus of everyday conversation, we found no significant differences in frequency of CRs and distribution of forms in the two task-oriented corpora, but many significant differences between CRs in task-oriented dialogue and everyday conversation. Our findings suggest that in task-oriented dialogues, humans use a cautious, but efficient strategy for clarification, preferring to present an hypothesis rather than ask the user to repeat or rephrase the problematic utterance. We also identified correlations between *function* and *form* features that can serve as a basis for generating more natural sounding CRs, which indicate a specific problem with understanding. In current work, we are studying data collected in a wizard-of-oz study in a multi-modal setting, in order to study clarification behavior in multi-modal dialogue.

Acknowledgements

The authors would like thank Kepa Rodriguez, Oliver Lemon, and David Reitter for help and discussion.

References

Christina L. Bennett and Alexander I. Rudnicky. 2002. The Carnegie Mellon Communicator Corpus. In *Proceedings of the International Conference of Spoken Language Processing (ICSLP02)*.

Lou Burnard. 2000. The British National Corpus Users Reference Guide. Technical report, Oxford Universiry Computing Services.

Herbert Clark. 1996. *Using Language*. Cambridge University Press.

Malte Gabsdil. 2003. Clarification in Spoken Dialogue Systems. *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving Ellipsis in Clarification. In *Proceedings of the 39th meeting of the Association for Computational Linguistics*.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Goteborg University.

Tim Paek and Eric Horvitz. 2000. Conversation as Action Under Uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*.

Tim Paek. 2003. Toward a Taxonomy of Communication Errors. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the Means for Clarification in Dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*.

Matthew Purver. 2004. CLARIE: The Clarification Engine. In *Proceedings of the Eighth Workshop on Formal Semantics and Dialogue*.

Verena Rieser. 2004. Fragmentary Clarifications on Several Levels for Robust Dialogue Systems. Master's thesis, School of Informatics, University of Edinburgh.

Kepa J. Rodriguez and David Schlangen. 2004. Form, Intonation and Function of Clarification Requests in German Task-orientaded Spoken Dialogues. In *Proceedings of the Eighth Workshop on Formal Semantics and Dialogue*.

David Schlangen. 2004. Causes and Strategies for Requestion Clarification in Dialogue. *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*.

Gabriel Skantze. 2003. Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

David R. Traum. 1999. Computational Models of Grounding in Collaborative Systems. In *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication*.