Form, Intonation and Function of Clarification Requests in German task-oriented spoken dialogues

Kepa Joseba Rodríguez and **David Schlangen**
University of Potsdam
Department of Linguistics / Applied Computational Linguistics
P.O. Box 601553
D-14415 Potsdam — Germany
{rodriguez|das}@ling.uni-potsdam.de

Abstract

We present a classification-scheme for describing the form (including intonation) and function of clarification requests (CRs) that is more fine-grained than extant classifications, and a study of a corpus of German task-oriented dialogues where we used this scheme to annotate the occuring CRs. Among the correlations between form and function we found was a hitherto undescribed correlation between intonation of CRs and their interpretation, which could possibly aid dialogue systems in interpreting CRs.

1 Introduction

Clarification requests (CRs), as exemplified by B's utterances in the mini-dialogues in (1), are of eminent theoretical as well as practical interest.

(1) a. A: Well, I've seen him.
B: Sorry, you *have* or you *haven't*?
b. A: Did you talk to Peter?
B: Peter Miller?
c. A: Did you bring a 3-5 torx?
B: What's that?

They are of theoretical interest because they are a prime example of a dialogue move that is concerned more with dialogue management than with conveying propositional information, and hence goes beyond what formal semantics was invented to model. Arguably even stronger is the practical interest in modelling CR, since practical dialogue systems are constantly confronted with situations where it would be beneficial if they could clarify their understanding of a user's utterance, or where they must interpret a clarification requested by the user.[1] (To give an impression of the frequency of this phenomenon even in human-human dialogue, in our corpus we found that around 5.8% of all turns were CRs.)

In this paper we hope to further both lines of inquiry, by offering a theoretically motivated and practically usable classification of CR uses and of CR forms, and by investigating the link between the two in a corpus of German spoken dialogues. While we replicate (for a different language) some of the results of earlier studies (Purver et al., 2001; Purver et al., 2003), we argue for, and show the use of, an analysis of form and function that is more fine-grained than that underlying those studies. We also make use of the fact that we had available information about intonation in our corpus—a feature that significantly influences the interpretation of CRs, as we show, and that could be used in practical dialogue systems to disambiguate CRs.

The remainder of this paper is organised as follows. In the next section we describe our multi-dimensional classification of form and function of

[1] The semantics of (some kinds of) CRs is modelled for example in (Ginzburg and Cooper, 2001; Larsson, 2003); see below for some remarks on the former analysis. There is a vast literature on dealing with clarifications in spoken dialogue systems, some very recent examples taking a more theoretical perspective include (Gabsdil, 2003; Larsson, 2003; Schlangen, 2004).

CRs and compare it to earlier classification attempts. In Section 3 we give details about the corpus study we conducted, whose results we present and discuss in Section 3.2. In particular, we discuss the links between form-features and function that are present in the corpus. We close with a discussion of the overall result, and of possible further work.

2 Classification of CRs

2.1 Earlier work

In a number of papers, the most recent of which is (Purver et al., 2001) (henceforth PGH), Jonathan Ginzburg and colleagues have put forward a scheme for classifying form and function of CRs, which we will now discuss and relate to the one proposed in this paper.

PGH classify CR-forms using the classes shown on the left in Table 1. While these classes achieve good coverage on the corpus (PGH report that only 0.5% of CRs were classified as `other`), we wanted to explore the influence of individual features of the form on the interpretation in more detail, and hence we further analysed these classes and devised a multi-dimensional classification. We will describe our schema in detail below, but to give an example of how it relates to PGH's, in our schema we 'factor out' the component "reprise" that is found in several of PGH's classes into a feature "relation to the antecedent" (`rel-antec`), which can take the values `repetition`, `reformulation` and `independent`, independently from other features. This allows us to make finer distinctions, for example between "Paris?" and "The capital of France?", which as a reply to "I'm going to Paris." would both be classified as `frg` by PGH. Our multi-dimensional approach also allows us to emphasise *similarities* between forms; for instance, PGH's classes `frg` and `lit` have in common in our approach the value for a certain feature (both are literal repetitions of material from the antecedent utterance), while having different values for other features. Using such fine-grained features, we can test for more fine-grained correlations between form and function.

While PGH's classification of CR-forms seems to be generally correct (just not as fine-grained as possible), their classification of CR functions (or *readings*, as they call them), shown on the right in Table 1, seems more problematic. In particular what they call the *clausal* reading of CRs seems to be difficult in practice to delineate from the other readings they define. For instance, given a situation as shown in (2), it is not clear why the *clausal* reading should not be able to play the function the authors assign to the *constituent* reading, namely to clarify a referent. (The other direction is more clearly distinguished: unlike the *clausal* reading, the *constituent* reading *cannot* clarify an acoustic problem.)

Class	Description	Example
non	Non-Reprise	*"What did you say?"*
wot	Conventional	*"Pardon?"*
frg	Reprise Fragment	*"Paris?"*
slu	Reprise Sluice	*"Where?"*
lit	Literal Reprise	*"You want to go to Paris?"*
sub	Wh-substituted Reprise	*"You want to go where?"*
gap	Gap	*"You want to go to ...?"*
fil	Gap Filler	*"... Paris?"*
oth	Other	Other

Class	Description	Paraphrase
cla	Clausal	"Are you asking/telling me that ...X..?"
con	Constituent	"What/who do you mean by 'X'?"
lex	Lexical	"Did you utter 'X'?"
corr	Correction	"Did you intend to utter X (instead of Y)?"
oth	Other	Other

Table 1: CR forms and readings as classified by (Purver *et al.* 2001)

(2) A: Did Bo leave?
B: Who?
clausal: For which x are you asking whether x left?
constituent: Who's Bo?

Moreover, it seems difficult to integrate CRs asking for clarification of intentions into this scheme:

(3) a. A: Push the red button.
B: Why?
b. A: Turn it on.
B: By pushing the red button?

To summarize, the problem seems to be that the readings defined by PGH still abstract over different reasons why one might want to make a CR—

they are still too close to the ambiguity of "what did you say?".[2] For these reasons we will in the next section propose a different classification of CR functions; first, however, we turn again to the *form* of CRs.

2.2 Surface form of CRs

We now go through the features we use to describe the form of the CRs.[3] A few selected examples for the different types are shown at the end of this section in (4).

Mood The possible values of the attribute `mood` are: a) *declarative*; canonical declarative word order or fragment without a verb with falling end-boundary tone.[4] b) *polar question*; fully realised syntactic polar interrogatives. c) *alternative question*; d) *wh-question*; e) *imperative*; f) *other*.

Completeness The possible values for the attribute `completeness` are: a) *particle*; or conventional phrase, e.g. "pardon?". b) *partial*; a syntactic fragment, normally a phrase. c) *complete*; a syntactically 'complete' sentence.

Relation to the antecedent The possible values for the attribute `rel-antec` are: a) *repetition*; parts of the problematic utterance are repeated *literally*. b) *addition*; something is added to a literal repetition (most often a *wh*-word). c) *reformulation*; a phrase is uttered that is co-referent to elements of the original utterance, but is not a literal repetition. d) *independent*; no elements of the problematic utterance are repeated or reformulated.

We also classify CRs according to the intonation with which they are uttered. Specifically, we look at the end-boundary tone, marking it use an encoding that is related to ToBI (Silverman et al., 1992), but somewhat simplified.

Boundary tone The values are: a) *rising* and b) *falling*, which correspond to (X)H% and (X)L%, respectively (X being an arbitrary tone).

A few examples for CRs of the types described above are shown below, with the classification according to the above scheme shown in `typewriter font`.

(4) a. K.: na hinten.
I.: vorne oder hinten?
K.: hinten.
(*K.: well, to the back – I.: to the front or to the back? – K.: to the back)*
`mood:alt-q,`
`completeness:partial,`
`rel-antec:addition,`
`bound-tone:falling`

b. I.: hm ist doch (ei)n Klacks für dich.
K.: hä?
(I.: hm, that shouldn't be a problem for you – K.: eh?)
`mood:other,`
`completeness:particle,`
`rel-antec:indep,`
`bound-tone:rising`

c. K.: ich hab(e) aber noch zwei Stäbe.
I.: du hast noch zwei Stäbe?
(K.: But I still have two bars. – I.: you still have to bars?)
`mood:decl,`
`completeness:complete,`
`rel-antec:repet,`
`bound-tone:rising`

d. I.: [. . .] und der grüne sitzt obendrauf.
K.: obendrauf?
(I.: [. . .] and the green one sits on top of it – K.: on top of it?)
`mood:decl,`
`completeness:partial,`
`rel-antec:repet,`
`bound-tone:falling`

2.3 Function of CRs

We also classify the function of each CR instance according to a multi-dimensional schema. The most important dimension is the one specifying

[2] These readings are realised technically by a straightforward formalisation of these paraphrases in an HPSG framework, using an *illocutionary-act* relation for the clausal reading and a relation *content* for the clausal readings, where both relations take signs as arguments. Since the formalisation is so close to the paraphrases (and is in any case not backed up by a formal semantics of the predicates used), we don't think we miss crucial details by using just the paraphrases in this discussion here.

[3] We initially also used word order as a classification feature, but since it turned out not to have any predictive power as to the possible function of a CR, we do not include it here.

[4] The name of this value is slightly misleading: it covers all cases of non-interrogative word order, i.e. both declarative sentences and fragments, and so a more appropriate (but less immediately understandable) name would be "non-interrogative".

	Level of action	Kind of problem	Example
1	execution / attention	channel	"huh?"
2	presentation / identification	Acoustic problem	"Pardon?"
3	signal / recognition	Lexical problem	"What's a double torx?"
		Parsing problems	"Did *you* have a telescope, or the man?"
		Reference resolution problem:	
		• NP-reference	"Which square?"
		• Deictic-reference	"Where is 'there'?"
		• Action-reference	"What's to kowtow?"
4	proposal / consideration	Problem with recognising or evaluating the intention	"Why?" "You want me to give you this?"

Table 2: Levels of action and associated problems

the likely source of the problem that lead to the need for clarification. This dimension is related to PGH's *readings*, but, as discussed above, needs to be more fine-grained and better defined. As the basis of our classification we use the well-known models of (Clark, 1996) and (Allwood, 1995), to which we add some further (sub-)levels. The other dimensions specify the *extent* and *severity* of the problem, as described below. Lastly, we also group under this heading a classification of the reaction to the CR.

Source of the problem The models of (Clark, 1996) and (Allwood, 1995) describe four levels of action involved in communication, each of which is a possible locus for communication problems. In Table 2 they are represented schematically, together with a specification of the kinds of problems that can occur on these levels, and some examples. As this specification shows, the levels can be further subclassified, and this we have done for our classification.[5]

The possible values for this feature correspond to the column "kind of problem" in the table. For reasons of space, we can only give the constructed examples in the last column of the table here.

Extent This feature describes whether the CR points out a problematic element in the problem utterance (e.g., "To Paris?", "I didn't hear the second word.") or not; its possible values are *yes* and *no*. Note that this is a *function*-feature, which may or may not be strongly connected to the *form*-feature "fragmental", but is logically independent, as the second example above, a full sentence that points out a problematic element, shows.

Expectation / Severity This dimension describes which action the CR initiator requests from the other dialogue participant, or, to look at it from another perspective, it describes how severe the problem was. The possible values are: a) *repetition/elaboration of previous material*; the CR initiator asks for a repetition/reformulation of material from the move to be clarified, possibly triggered by a complete understanding failure. b) *confirmation of the hypothesis*; the CR initiator asks for a confirmation of her/his understanding about the content of the move to be clarified. I.e., a hypothesis could be drawn, but agent is not confident about its correctness. (4-b) above is an example for the former, (4-a) for the latter.

Reply to the CR This feature classifies the reply to the CR, not the CR itself. Its possible values are: a) *y/n-answer*; b) *repetition*; an answer that repeats an element of the problem utterance literally. c) *reformulation*; an anwer that reformulates an element. d) *elaboration*; an answer that elaborates on (an element of) the problem utterance, adding information. e) *word definition*; an answer to a lexical question ("what does x mean?"). f) *no reaction*; the CR addressee did not react.

Satisfaction of the CR-initiator This feature records the reaction of the CR initiator to the reaction of the CR addressee. The possible values are: a) *happy*; the CR initiator seems satisfied with the reply; this can be taken as an indication that the interpretation of the CR addressee was correct. b) *unhappy*; CR initiator renews request for clarification.

[5] (Gabsdil, 2003) and (Larsson, 2003) similarly use these models to classify CRs, and they are roughly at the same level of fine-grainedness. (Schlangen, 2004) uses a more fine-grained classification that is motivated by a formal semantic / pragmatic processing model, but to strike a balance between detailed analysis of the phenomenon and making annotation possible, we have decided on the fewer levels described here.

3 The Corpus study

3.1 Material and Method

Material We used the Bielefeld Corpus of German task-oriented human-human dialogue (SFB-360, 2000) (the scenario is that one dialogue participant (DP) gives instructions to the other DP to build a model plane), which consists of 22 dialogues, with 3962 dialogue turns and 35813 words.

Method In a first step, we identified the turns containing CRs, which we then annotated for form and function, using the MMAX-tool (Müller and Strube, 2001). Annotation of the form classification features was straightforward, as their values can easily be read off of the surface form of the CR, or, in the case of `rel-antec`, from CR and problem utterance. The *function* of a CR of course cannot as easily be seen from the form—to find whether there is a reliable link is one goal of the present study, after all. We used the reply of the CR addressee, and the reaction of the CR initiator to that reply as a guide for the interpretation that was chosen by the DPs. Hence what we annotated as 'function' could more properly be called "mutually agreed upon interpretation of the CR"—and that is not necessarily what the CR initiator might initially have had in mind. Since "overanswering" in certain configurations systematically addresses several different problem sources (for example, a reformulation of content answers both acoustic understanding problems as well as reference resolution problems), this is a real methodological problem for finding a link between form and problem source. We circumvented this problem by defining ambiguity classes for use in the cases where we could not make a decision; this weakens the overall correlations we report below, but makes the ones we did find between form features and unambiguously identified functions more valid.[6]

[6]This strategy is more cautious than the one chosen by PGH. As they say, in cases of ambiguity "the response(s) of the other DPs were examined to determine which reading was chosen by them. The ensuing reaction of the CR initiator was then used to judge whether this interpretation was acceptable." However, this method is not infallible, as their own example shows:

(i) George: [...] with a piece of spunyarn in the wire.
Anon1: Spunyarn?
George: Spunyarn, yes.
Anon1: What's spunyarn?
George: Well that's like er tarred rope.

PGH use this as an example where the original interpretation was incorrect; however, in our opinion an interpretation seems equally likely where Anon1 first wanted to clarify acoustic understanding, and, once this was accomplished, clarified lexical understanding. To be on the safe side, we annotated such cases with superclasses combining the subclasses it is ambiguous in.

```
FORM:
  distance: { 1 | 2 | 3 | 4 | 5 | more }
  mood: { none | decl | polar-q | wh-q | alt-q |
          imp | other }
  form: { none | particle | partial | complete }
  rel-antec: { none | addition | repet |
               reformul | indep }
  boundary-tone: { none | rising | falling | no-appl }
FUNCTION:
  src: { none | acous | lex | parsing | np-ref |
         deictic-ref | act-ref | int+eval | src-3 |
         src-2+3 | src-2+4 | src-3+4 | src-all }
  extent: { none | yes | no }
  severity: { none | cont-conf | cont-rep | no-react }
  answer: { none | ans-repet | ans-y/n | ans-elab |
            ans-reformul | ans-w-defin | no-react }
  happiness: { none | happy-yes | happy-no |
               happy-ambig }
```

Figure 1: Annotation scheme

Annotation scheme The annotation schema basically just implements the distinctions described in Section 2, with ambiguity classes for function as discussed above. It is shown in Fig. 1. Note that we also recorded the distance between the CR and the problem utterance.

3.2 Results

We identified 230 CRs in the 3962 turns we looked at; this indicates that with 5.8% of turns this is a rather frequent phenomenon in our corpus. (PGH: just under 4%, but their corpus contained general conversation, which might account for the difference.) The results of classifying these instances and of testing for dependence between features are reported in this section.

3.2.1 Distribution

Clarification seems to be a very local phenomenon: in our corpus, 95% of all clarifications target the immediately preceding utterance (PGH: 85%). This high number might reflect the

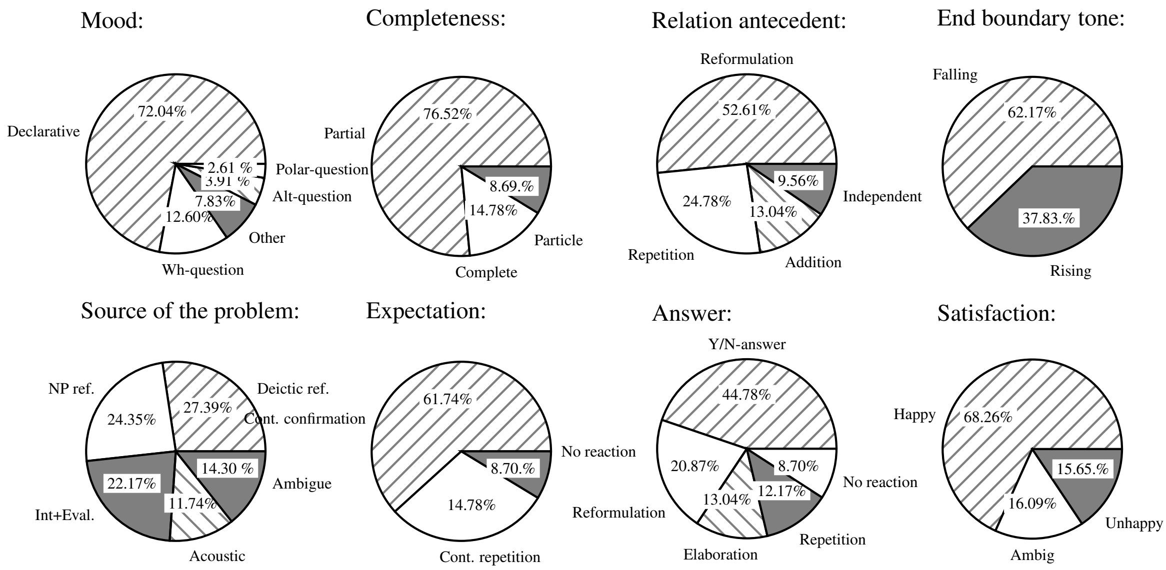

Figure 2: Distribution of values for form-features (top) and function-features (bottom)

task-oriented nature of the corpus, where grounding presumably is more cautious compared to free conversation, and potential problems are clarified immediately.

Distribution of forms The frequencies for values of the form-features are given by the pie-charts in the top row of Figure 2. As this figure shows, the overwhelming majority (76.5%) of the CRs in our corpus were fragmental in form (PGH: 42.4%). Since we separated our analysis into several dimensions, we can further analyse the class of fragmental CRs: 62.6% of them were reformulations of previous content, 24.8% were repetitions. Another distinction not made in earlier studies is that between rising and falling intonation. Using these features, we can access different (sub-)types of what PGH collectively call "reprise fragments". Indeed, four of the five most frequent types of CRs were classified as syntactically partial (i.e., the value for `completeness` is `partial`), with either `falling` or `rising` as value for `bound-tone` and either `repet` or `reform` as value for `rel-antec`. (The one other type in the "top-five" being that of conventional CRs.) We come back to these distinctions when we report the correlations between form and function we found.

Another interesting observation is that most CRs take up material from the problematic utterance in some form, with only 9.6% of CRs being fully independently formulated. Overall, these numbers seem to confirm the findings of PGH regarding distribution of forms, showing that at least for speakers of English and German behaviour with respect to clarification seems comparable—useful to know for designers of dialogue managers for multi-lingual dialogue systems.

Distribution of functions The distribution of values for the function-features can be seen in the bottom row of Figure 2 (with the exception of the feature `extent`, whose two values `yes` and `no` were chosen 87.8% and 12.2% times, respectively).

As this figure shows, the most frequent problems were related to resolving references (just above 50%, with 27.4% clarifying deictic references, and 24.4% clarifying NP reference). 14.3% of the CRs were annotated with a super-class, meaning that their function was ambiguous in the context. However, most instances, and most types, could be classified unambiguously. The distribution of super-classes is instructive, showing for example that the different kinds of problems at level 3 could be distinguished fairly well.

We found only one instance of a lexical problem, making our corpus non-representative for this type of CR. We speculate that the reason is that the vocabulary in this domain is very restricted and domain specific, and known to the DPs. This might also explain the relatively low frequency of acoustic problems, since a restricted vocabulary may make recognition easier.

Most CRs in our corpus point out a specific element in the problem utterance—this of course correlates with the high number of fragmental CRs found. Only 8.7% of CRs in our corpus failed to elicit a response ((Purver et al., 2003): 17%); again, this seems to be a difference between task-oriented dialogue, where the task demands that problems be clarified, and free conversation.

3.2.2 Correlations

We used χ^2 to test for (in)dependence between features of the surface form and function of CRs (we used Yates' correction to account for cases where due to data sparseness there were expected values below 5), and if there was a significant dependence, Pearson's ϕ to determine the strenght of the correlation. The results of this test are shown in Table 3, where the rows are the form dimensions and the columns those of the function, and the cells show the results of testing for (in)dependence between these variables (showing χ^2, χ^2 with Yates' correction, and Pearson's ϕ). Note that all tests are significant at P=0.001. For reasons of space we can only pick out the most relevant findings here for further discussion.

One very interesting result is that intonation seems to disambiguate fairly reliably between CRs clarifying reference and those clarifying acoustic understanding, with rising boundary tones being significantly more often used to clarify acoustic problems and less often than expected to clarify reference resolution problems, and complementary correlations for falling tones. (The confusion matrix is shown in Table 4.) A similar distinguishing tendency is shown by reformulations vs. repetitions, with the former being significantly often NP reference resolution questions and the latter acoustic clarifications.

Looking at `mood` vs. `answer`, one can see that declaratives in general prompt yes/no-answers (and hence confirmations of hypotheses) more than reformulations of content, which in turn is the most likely reaction to *wh*-questions. These are nice results, showing that despite the fact that both readings are in principle available for fragments (cf. PGH), more clarity is achieved if a di-

	`rising`	`falling`	
`int+eval`	24 (21.18)	32 (34.82)	56
`deictic-ref`	**8** (20.43)	**46** (33.57)	54
`np-ref`	**8** (18.91)	**42** (31.09)	50
`acous`	**23** (9.08)	**1** (14.92)	24
`src-2+3`	3 (2.27)	3 (3.73)	6
`src-2+4`	**11** (6.81)	**7** (11.19)	18
`src-all`	**1** (0.76)	1 (1.24)	2
`lex`	0 (0.38)	**1** (0.62)	1
`src-3+4`	9 (7.19)	10 (11.81)	19
	87	143	230

χ^2 Total: 63.23 (YC: 56.59); df = 8; $\phi = 0.52$

Table 4: `src x bound-tone`

alogue system for example produces such forms only if it wants to get a hypothesis confirmed, and *wh*-questions if it needs more information about an element of the problem utterance. Moreover, if the hypothesis is one about the referent of an NP, a reformulation is the best bet; if it is one about acoustic understanding, a literal repetition might be better.

3.3 Reliability

Although the complete annotation was only performed once (by one of the authors), we did test for reliability of what is intuitively the most problematic feature, namely `source of the problem`. This feature was annotated by a second annotator, resulting in a κ (Carletta, 1996) of 0.70. While this is not great (values between .67 and .8 are often seen to allow only tentative conclusions), it is comparable to the results reported by PGH (0.75), and reflects the difficulty of the task.

Where we cannot report reliability yet is for the task of identifying CRs in the first place. This is not a trivial problem, which we will address in future work.[7]

4 Summary and Further Work

We have presented a fine-grained classification scheme for form and function of clarification requests. This scheme was used to annotate a corpus of task-oriented dialogues, where about 4% of all turns were found to be CRs—this confirms the observation that clarification is a quite frequent phenomenon. Our fine-grained annotation scheme,

[7]As far as we can see, PGH have not tested for reliability of doing this task either.

	source	severity	extent	answer
mood	*indep.*	χ^2-Σ: 106.52/96.58; df = 8; $\phi = 0.48$	χ^2-Σ: 112.04/101.31; df = 4; $\phi = 0.70$	χ^2-Σ: 72.90/72.64; df = 20; $\phi = 0.28$
bound-tone	χ^2-Σ: 63.23/56.59; df = 8; $\phi = 0.52$	*indep.*	χ^2-Σ: 14.85/13.29; df = 1; $\phi = 0.25$	*indep.*
rel-antec	χ^2-Σ: 142.85/114.62; df = 24; $\phi = 0.46$	χ^2-Σ: 66.16/59.87; df = 6; $\phi = 0.38$	χ^2-Σ: 98.49/90.55; df = 3; $\phi = 0.65$	*indep.*
completeness	*indep.*	χ^2-Σ: 35.88/31.50; df = 4; $\phi = 0.28$	χ^2-Σ: 94.54/86.98; df = 2; $\phi = 0.64$	*indep.*

Table 3: χ^2 values for combinations of form- and function-features

and the fact that we annotated intonation, allowed us to find correlations that have hitherto been unnoticed, such as that described above between intonation of CRs and their relation to the antecedent utterance (repetition or reformulation) on the one hand and reference resolution function or acoustic clarification on the other hand. Information like this could be of much use in dialogue systems that are faced with the task of interpreting CRs by the user which in theory are often multiply ambigous.

In further work we plan to connect our findings to general theories of the interpretation of intonation in discourse (e.g. (Gunlogson, 2001)), and we also plan to collect more data, with which then automatic classifiers could be trained. Another interesting extension of the research presented here would be to also annotate features such as "quality of the communication channel", or "frequency of clarified word", which could further aid interpretation.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

References

Jens Allwood. 1995. An activity based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics 76, Göteborg University, Göteborg, Sweden.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge, UK.

Malte Gabsdil. 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, USA.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *Proceedings of the 39th Meeting of the ACL*, Tolouse, France.

Christine Gunlogson. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California, Santa Cruz, CaliforniaUSA, December.

Staffan Larsson. 2003. Interactive Communication Management in an Issue-based Dialogue System. In Ivana Kruijff-Korvayová and Claudia Kosny, editors, *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue. Diabruck*, pages 75 – 82, Saarbrücken, Germany, September. Universität des Saarlandes.

Christoph Müller and Michael Strube. 2001. MMAX: A Tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, Seattle, USA, August.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the Means for Clarification in Dialogue. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial01)*, pages 116–125, Aalborg, Denmark, September. Association for Computational Linguistics.

Matthew Purver, Patrick G.T. Healey, James King, Jonathan Ginzburg, and Greg J. Mills. 2003. Answering clarification questions. In Alexander Rudnicky, editor, *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue(SIGdial03)*, Sapporo, Japan, July.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue (SIGdial04)*, Boston, USA, April.

SFB-360. 2000. Bielefeld Corpus. URL: `http://www.sfb360.uni-bielefeld.de`.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, October.