

Recovering from Non-Understanding Errors in a Conversational Dialogue System

Matthew Henderson
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge, CB2 1PZ
mh521@cam.ac.uk

Colin Matheson
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
colin@inf.ed.ac.uk

Jon Oberlander
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
jon@inf.ed.ac.uk

Abstract

Spoken dialogue systems can encounter different types of errors, including non-understanding errors where the system recognises that the user has spoken, but does not understand the utterance. Strategies for dealing with this kind of error have been proposed and tested in the context of goal-driven dialogue systems, for example by Bohus with a system which helps reserve conference rooms (Bohus and Rudnicky, 2005). However there has been little work on possible strategies in more conversational settings where the dialogue has more open-ended intentions. This paper looks at recovery from non-understanding errors in the context of a robot tourguide, and tests the strategies in a user trial. The results suggest that it is beneficial for user enjoyment to use strategies which attempt to move the dialogue on, rather than getting caught up in the error by asking users to repeat themselves.

1 Introduction

The handling of understanding errors is an important consideration in the design of a spoken dialogue system. Many dialogues take place in difficult conditions, with spontaneous speech, large vocabularies, varied user populations and uncertain line quality (Bohus and Rudnicky, 2005). These conditions make understanding errors very likely during the course of a dialogue.

There are two types of understanding error which a spoken dialogue system may encounter: *non-understandings* and *misunderstandings*. A non-understanding is where the system fails to extract a valid semantic representation of what

the user said. A misunderstanding is where a valid representation is extracted which happens to be incorrect. While detecting misunderstandings requires some thought, non-understanding errors are immediately apparent to the system due to the failure of the natural language understanding component.

This paper looks at strategies for dealing with non-understanding errors in the context of conversational spoken dialogue systems, as opposed to slot-filling or more generally goal-driven approaches. In such goal-driven systems, the user and the system typically work together to accomplish a specific task, for example booking a flight, finding a restaurant or reserving a conference room. This normally involves the system obtaining some information from the user (or filling a list of slots with their values), checking a database, and then completing the task. In a more conversational dialogue system the only real task is to take part in an interaction which is interesting and enjoyable for the user, although in work related to ILEX (Mellish et al., 1998) the system may have the loose goal of communicating prioritised pieces of information, and the research reported here is in this tradition. There may not be a definitive distinction to be drawn between what we have termed ‘conversational’ and ‘goal-driven’ systems, apart from pointing to the typical need that the latter have to fill slots with information elicited from the user, while there is no such target in the former.

The INDIGO project (Vogiatzis et al., 2008), (Konstantopoulos et al., 2009) followed the ILEX notion of opportunistic language generation, adapting the approach to spoken interactions with a robot museum guide. A later version of

the guide was tested with an initial ‘fake’ strategy (see below) for avoiding repetitions of the standard ‘*Could you please repeat that*’ form for dealing with non-understandings. This unreported pilot work is updated here and extended to include a set of non-understanding error recovery strategies which aim to improve user enjoyment of conversational dialogues with a robot tourguide. The strategies are tested in a user trial which is designed to elicit answers to the following questions:

- Can user satisfaction be increased by using smart strategies to deal with non-understanding errors in a conversational dialogue system?
- How does the use of such strategies affect the user’s perception of the dialogue and the dialogue system?
- How do the strategies compare to each other in terms of user satisfaction, and in particular is it important to employ a variety of strategies?

2 The Tourguide Dialogue System

The Tourguide Dialogue system was built in order to investigate non-understanding error recovery strategies in a conversational domain. The chosen application is that of acting as a tourguide in an exhibition. The dialogues consist of the system describing an item, and then taking questions from the user. Specifically, the system talks about 3 items which can be found in the Informatics Forum at the University of Edinburgh.

During the course of the tourguide dialogues, the point where most errors are anticipated is when the system asks ‘Do you have any questions about this?’ This obviously constitutes an extremely open question, and the lack of constraints on the user’s input results in a high probability of a non-understanding occurring. At all the other points where the system elicits input from the user it has full initiative and can supply the speech recognition module with a set of highly constrained expectations, whereas in the situation above (although the system does attempt to predict the input), the range of possibilities is very large. This is thus a good context in which to investigate strategies for dealing with non-understanding errors. The system is designed to be programmed with a library of error recovery strategies. For a list of the strategies implemented,

see Table 1.

Table 1: Summary of Strategies

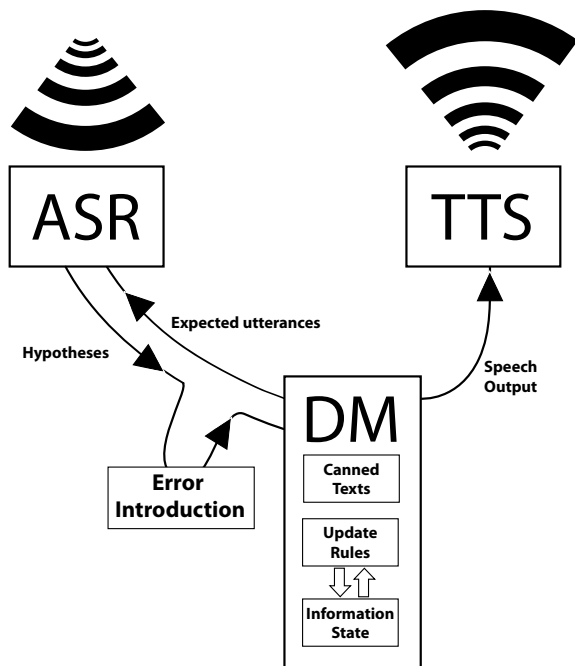
Name	Description
Subsume	Ask if the user would like to hear more information about the item E.g. : “Would you like to hear a little more about Paolozzi and his sculptures?”
Subsume Split	Ask if the user is more interested in hearing about aspect A or aspect B of the item E.g. : “Well, are you more interested in finding out more about Paolozzi himself, or his sculptures?”
Fake	Fake having forgotten to say something of interest about the item E.g. : “I meant to add; one of Paolozzi’s most famous works can be found here in Edinburgh. At the top of Leith Walk, there are sculptures of human body parts, including a giant foot, by Paolozzi.”
Please Repeat	Ask the user to repeat their question E.g. : “Please could you repeat that? Just say ‘no’ if you have no more questions.”

The dialogue manager is implemented in Prolog, using the Trindikit framework. It is a hand-crafted dialogue system, which uses the information state model to hold the system’s beliefs and a set of update rules which define the system’s actions. See Larsson and Traum (2000) for a summary of information state and dialogue management using the framework.

For speech input and output, modules developed by Acapela Group¹ for the INDIGO project were used. As mentioned, at each point in the dialogue at which user input is expected, a hand-crafted list of possible user utterances is sent to the speech recogniser. For example, at a point

¹<http://www.acapela-group.com>

Figure 1: Overview of the Tourguide Dialogue System



where the system asks for questions on a particular exhibit, the speech recogniser is supplied with a list of questions which were predicted by the system designer. Language generation uses only simple templates, which are sent to the Acapela text-to-speech component.

In order to ensure that non-understanding errors occur at a consistent and non-negligible rate, a component is introduced into the system between the speech recogniser and the dialogue manager which serves to introduce errors at a predetermined rate. If a real non-understanding hasn't happened after 3 questions, an error is introduced by throwing away the speech recognition result and simulating a null user act. An overview of the structure of the system is presented in Figure 1.

3 Non Understanding Error Strategies

3.1 Motivating the Strategies

There have been studies using human communication to investigate how human agents deal with non-understanding errors, in the hope that this can be applied to spoken dialogue systems (Zollo, 1999; Skantze, 2003; Koulouri and Lauria, 2009). Wizard of Oz methods allowed analysis of dialogues between human users linked by computer

systems. To emulate a real spoken dialogue system, the Wizard sees the output of a speech recogniser, and the user either listens to the output from a speech synthesiser or a vocoder. All three of these studies focused on relatively restricted goal-driven dialogues, where the user and system had to work together to accomplish a task with a clear target outcome.

In these experiments, the wizards and users were always naive participants, and the responses of the wizards were not limited in any way (with the exception of some of the conditions in Koulouri and Lauria (2009)). This allowed the experimenters to analyse how a human might try to deal with speech recognition errors when trying to conduct a dialogue.

A common theme in all three studies was the importance of using error recovery strategies which help the dialogue to progress. It was found that wizards will often ask task-related questions, the answer to which subsumes the information missed by the non-understanding. The example question from Skantze (2003) below illustrates this:

wizard

Do you see a wooden house in front of you?

user

Yes crossing address now. (*Actually: I'm passing the wooden house now.*)

wizard

Can you see a restaurant sign?

Here the Wizard asks a follow-up question which is related, in that its answer implies the information they just missed. Skantze found that this strategy not only improved the understanding of the following utterances, but also resulted in higher user perception of task success.

Other wizard of Oz studies have looked specifically at evaluating error recovery strategies (Schlangen and Fernández, 2006; Rieser et al., 2005), and Bohus implemented a variety of non-understanding error recovery strategies in a real dialogue system (Bohus, 2007); relevant findings are summarised in Bohus and Rudnicky (2005). Again this study focuses on a goal-driven dialogue system, specifically a system which helps users book conference rooms. In the current context, one of the most interesting strategies implemented was called *MoveOn*, where the system

would continue by asking a new question when faced with a non-understanding. An example is:

Sorry, I didn't catch that. One choice would be Wean Hall 7220. This room can accommodate 20 people and has a whiteboard and a projector. Would you like a reservation for this room?

This strategy performed well with respect to recovery rate, i.e. how often the following user response was correctly understood. Bohus and Rudnicky explained its success by comparing it to other strategies, which would generally ask the user to repeat themselves, or rephrase their answer. In those cases it is unlikely that the system will be able understand the user's intention as it did not understand the input the first time. This process is prone to turning into a spiral of errors, with the user getting more and more frustrated. Frustration can affect the user's voice, in turn adversely affecting the Automatic Speech Recognition. On the other hand, with *MoveOn*, the system abandons the current question and tries a new line of attack.

The *MoveOn* strategy is related to the recommendations of Zollo, Skantze and Koulouri, and it seems from these studies that the idea of moving on, and asking a new question can be very effective. However it is not entirely clear how this strategy can be adapted to use in a conversational dialogue system. It is to this question that we now turn.

3.2 The Strategies

In a conversational dialogue system, there is as noted above no real goal in the sense of information to be elicited and acted upon, so it is not clear what constitutes a 'task-related' question in the sense used in the above studies. Indeed, in the Tourguide dialogue system, it is not usually the robot which is asking questions of the user but the other way around.

The general aim is thus to progress the dialogue smoothly when the user has just asked a question about an item in the exhibit which the system hasn't been able to understand. The first strategy which attempts to do this is called *Subsume* (see Table 1 for a summary of all the strategies, with examples). The *Subsume* strategy asks if the user is interested in finding out more about the item, it then waits for a response – any response –

and then proceeds to output a short text about the item. The text is designed to incorporate answers to a lot of the possible questions which the user may have asked. The strategy tries to broaden the user's goal from obtaining a specific piece of information to just hearing some general interesting information about the piece.

The second strategy is *Subsume Split*, which is similar to *Subsume* but gives the user a choice of what subsuming information they prefer. The questions for every item in an exhibit should broadly be able to be split into two categories. For example, for an artefact like a sculpture, these could be (a) questions about the artefact's creator and (b) questions about the artefact and other examples of the creator's work. In giving the user a binary choice the hope is that the information subsequently presented will be of more interest, and more closely related to their original question. Whereas *Subsume* did not rely on the next utterance being understood, *Subsume Split* requires the speech recogniser to distinguish between two possible answers. This is of course back in line with standard system-initiative approaches in which speech errors are much less of a problem as speech recognition generally works well in constrained contexts.

The last strategy implemented is called *Fake*, an approach which was investigated in the non-published pilot study of the INDIGO project mentioned above. The idea is for the system to pretend it has forgotten to include a piece of information. Although it is clearly unrealistic for a robot to be forgetful, it is hoped that the piece of information which the system provides is sufficiently interesting to justify saying it instead of answering the user's actual question. In the pilot study, the system returned to the original dialogue state after the diversion, asking Now, did you have a question?, and attempted to combine the speech recogniser scores from the two user turns to gain a higher confidence hypothesis. The current version implements the same dialogue moves but does not combine the recognition scores.

As well as these strategies, the system has a default *Please Repeat* strategy which is used as a baseline. This is used as a backup strategy whenever the other strategies are exhausted in a particular dialogue.

An excerpt from a typical dialogue follows, where an error is introduced and the system uses

Table 2: Questionnaire

Communication	Agent	Attitude
The system understood what I said	The system is intelligent	I enjoyed talking to the system
My conversation with the system flowed smoothly	The system was helpful	I felt confused when talking to the system*
It was clear what was happening when the system did not understand me	My conversation with the system was interesting	I felt frustrated when talking to the system*

For further analysis, responses to questions marked with an asterisk are converted from (1, 2, 3, 4, 5) to (5, 4, 3, 2, 1) so that higher numbers correspond to higher user satisfaction as with the other questions.

the SubsumeSplit strategy.

Excerpt from Typical Dialogue

System: This is a sculpture by Eduardo Paolozzi, a sculptor from Edinburgh. Paolozzi was interested in science, and our quest to understand more about the Universe and ourselves. The themes of his work relate to the research done here in Edinburgh. This piece is different in style to usual Paolozzi sculptures in its simplicity and smoothness. Do you have any questions?

User: Umm. Where is he from?

System: Paolozzi was born in Leith, in north Edinburgh. Do you have any other questions?

User: Why did he chose this shape?

System: I'm not sure why the sculpture is this shape. You'd have to ask Paolozzi himself! Do you have any more questions?

User: What does it represent?

-Non-understanding error introduced-

System: Well, are you more interested in Paolozzi himself, or his sculptures?

User: Eh, his sculptures.

System: Paolozzi's sculpture work was inspired by some of the great minds in the history of science. He invented a method of casting bronze sculptures, analogous to the process of creating collages. His early collage work in Paris pre-empted Pop Art. There are 4 of his sculptures in the Informatics Forum, and his work can be seen throughout the cities of Britain. Do you have any more questions?

4 Experimental Setup

The experiment consists of running a user trial with the system in 3 different modes. In the first mode (*mixed*) the system uses all the strategies shown in Table 1. In the second (*single*) the system uses a single strategy, either Subsume Split, or Fake. And in the last (*pleaserepeat*) it has no strategies except the default Please Repeat.

Participants were sat in front of a laptop running the Tourguide Dialogue System and asked to conduct a 10 to 15 minute long dialogue. The laptop screen displays the text as it is synthesised by the system, and also a list of example responses at each stage in the dialogue. Other than the length requirement, the users were not given any particular tasks to achieve in order to approximate a natural interaction with a conversational system. The participants were shown pictures of the three items which the system can talk about, and were told to ask the questions they believe they might ask if they were actually at the exhibition with a robot.

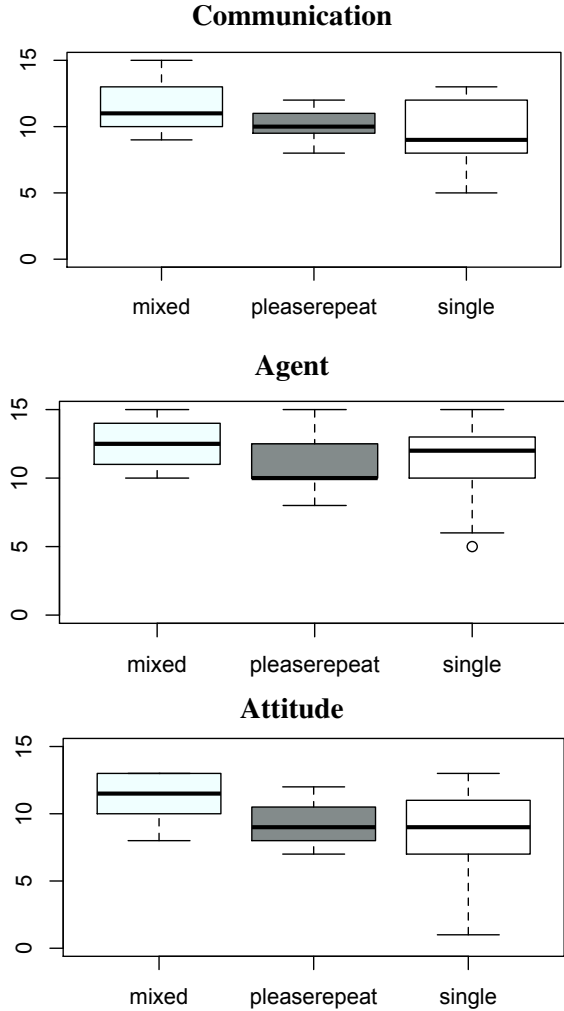
As mentioned, the system is configured to introduce a non-understanding error at every third question asked, as long as a real error did not occur in the previous three turns. The error rate is thus relatively consistent across the dialogues. The misunderstanding error rate due to incorrect speech recognition on user questions was 18%, this did not change significantly between conditions.

At the end of the interaction, participants were asked to fill in a questionnaire which includes a series of statements with which the participant must specify their level of agreement on a scale of 1 to 5. These statements are listed in Table 2, and are designed to measure the user's satisfaction along multiple dimensions. These questions serve to quantify the quality of the dialogue from the user's perspective better than an objective score such as dialogue length could estimate.

5 Results

Data from 58 participants in total was gathered, 14 in the *mixed* condition, 29 in the *single* condition (14 with Fake strategy and 15 with Subsume Split) and 15 in *pleaserepeat*. The questions on the questionnaire are grouped into three collections as shown in Table 2. The col-

Figure 2: Breakdown of Questionnaire Results



lections correspond respectively to the quality of the *Communication*, the user’s perception of the system as an *Agent*, and the user’s *Attitude* towards the dialogue. Within the three collections, the question answers are found to be highly correlated. The individual scores of each question in a collection are combined by simply adding them together, giving a collection score between 0 and 15.

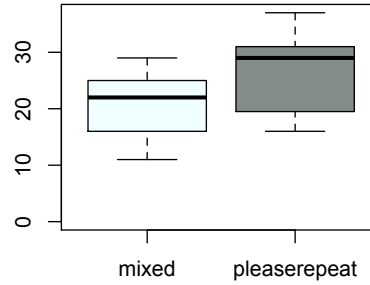
Figure 2 shows the results for each collection in each of the 4 conditions as box-whisker plots. In this paper, values more than $3/2$ times the interquartile range lower than the first quartile are treated as outliers, as are values that are analogously higher than the third quartile.

6 Analysis of Results

6.1 Analysis of Questionnaire Data

Kruskal-Wallis tests are used to test the hypothesis that the boxplots shown in Figure 2 represent

Figure 3: Number of Questions Asked



distinct distributions, i.e. that there is some difference in the distribution of a collection score between the groups. These tests suggest further investigation into *Communication* and *Attitude* (with the probability of the null hypothesis being less than 0.02), but not into the *Agent* scores.

Pairwise Mann-Whitney tests in the *Communication* and *Attitude* data are performed to test whether the differences between the pairs of groups are significant. Bonferroni correction is used to account for the fact that there are 3 comparisons for each collection, so a threshold of 0.015 ($< 0.05/3$) on the p -value is chosen. From this analysis, the following comparisons are found to be significant:

- **Communication**
mixed > pleaserepeat;
mixed > single.
- **Attitude**
mixed > single.

These results imply that the quality of the *Communication* of a dialogue (recall a combination of the flow of conversation, clarity of the system’s actions and how well it seems the system understands the user) is significantly improved by using a mixture of error recovery strategies as against a single strategy, as well as against the baseline please repeat. Variety in the dialogue may give the user an impression of a richer dialogue. Figure 3 shows how the number of questions asked is much higher in the *pleaserepeat* condition than in *mixed*.

This is because the mixture of strategies allow the system to do more of the talking, and to answer many of the user’s questions before they are asked. Less user questions means less possibility for error, and thus better dialogues. The strategies are exploiting the fact that users don’t mind

being provided more information than they originally asked for.

As mentioned, the Attitude measure is a combination of user enjoyment and lack of confusion and frustration. This is found to be significantly better in the `mixed` condition than in the `single` condition, but the comparison between `mixed` and `pleaserepeat` is not statistically significant ($p = 0.05$).

Note that the `mixed` condition is at an advantage relative to the `single` condition because it will take longer before the system resorts to the Please Repeat strategy. Therefore in the comparisons we must bear in mind that there are on average more Please Repeats being issued in the `single` condition.

6.2 Discussion

The strategies effectively use errors as an opportunity to tell the user something which it believes could be of interest. In a more complex system, the information provided could be tailored using a user model, as in the approach noted in the Introduction (Mellish et al., 1998). It is worth noting that if a system can opportunistically exploit errors to actively improve user experience, it could weaken the typical inverse correlation between user satisfaction and non-understanding rate, or at least, the rate of repetition-requests (Walker et al., 2000). Demonstrating this remains a matter for future work, however, since the current study specifically maintained a constant non-understanding rate across conditions, rather than treating it as an independent variable.

Lastly, it is interesting to investigate some of the correlations between the individual questionnaire answers using Pearson's correlation tests. The users' enjoyment is not correlated with how clearly they understand what the system is doing when an error occurs. This implies that it is not necessarily important for the user to understand what motivates the system's dialogue turns for them to enjoy the interaction. This appears to contradict the findings of Hockey et al. (2003) among others, which show that making the system 'visible' to the user increases the level of task success. The suggestion is therefore that the latter finding only applies in goal-driven dialogue systems, and so although the user must have some idea of what is motivating the system, it is not necessarily as important in more conversational settings.

7 Conclusions

In summary, this study has provided evidence that these new strategies, which use the idea of moving the dialogue on when the system has little or no input from the user, can have a positive effect on overall user satisfaction. It is shown that the benefit of such strategies is in using them as a strategy, and giving a conversational dialogue system a variety of error handling techniques.

Use of all of the strategies was significantly beneficial for the dialogues in the three dimensions measured in the questionnaire. Therefore, when designing a conversational dialogue system, it is worthwhile putting thought into the design of error recovery strategies which are more complex than asking the user to repeat or rephrase themselves. It is particularly beneficial to ensure that there is a variety of strategies available to the system, both to increase the variation in the dialogue and to make the individual strategies more effective. This has been confirmed experimentally in the goal-driven domains (see Section 3.1), and this paper provides initial supporting evidence in conversational, less goal-directed applications.

8 Future Work

A number of potential further investigations are possible:

- Presumably user enjoyment in a conversational dialogue system tends to degrade as error rates increase (Walker et al., 2000). It would be interesting to compare how quickly this degradation occurs when different error recovery strategies are employed. It is possible that strategies such as those presented here would help to maintain a minimal level of enjoyment longer.
- At one end of the spectrum, some goal-driven dialogue systems can be associated with a single objective metric of task success, independent of user impressions. Towards the other end of that spectrum, conversational systems like museum tour guides should allow different visitors to pursue distinct tasks, or single visitors to shift from one task to the other, and even interleave them. In such cases, more work is needed to identify the varying criteria for success for any given user.
- In this study the `mixed` condition chooses

strategies at random. It might be useful to investigate whether there exists a better-than-random policy. Bohus et al. have looked at this question in goal-driven applications (Bohus and Rudnicky, 2005; Bohus et al., 2006).

- More strategies could be investigated, possibly ones which exploit a user model to select pieces of information to impart. The current strategies use text which is the same for all users, whereas the use of a full language generation system producing dynamic texts would not only allow for tailoring to the user but also cause the strategies to be used more than once in a given part of the dialogue.

Acknowledgements

The work reported here was supported by both the INDIGO (IST-045388) and Help4Mood (ICT-248765) projects. We are grateful to Vasilis Karaiskos for assistance in completing the evaluations.

References

- Dan Bohus and Alex Rudnicky. 2005. Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.
- Dan Bohus, Brian Langner, Antoine Raux, Alan W. Black, Maxine Eskenazi, and Alex Rudnicky. 2006. Online supervised learning of non-understanding recovery policies. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 170–173, dec.
- Dan Bohus. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. thesis, Carnegie Mellon University.
- Beth A. Hockey, Oliver Lemon, Ellen Campana, Laura Hiatt, Gregory Aist, James Hieronymus, Alexander Gruenstein, and John Dowding. 2003. Targeted Help for Spoken Dialogue Systems: intelligent feedback improves naive users' performance. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 147–154. Association for Computational Linguistics.
- Stasinios Konstantopoulos, Athanasios Tegos, Dimitris Bilidas, Ion Androutsopoulos, Gerasimos Lampouras, Prodromos Malakasiotis, Colin Matheson, and Olivier Deroo. 2009. Adaptive natural language interaction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 37–40, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Theodora Koulouri and Stasha Lauria. 2009. A WOz framework for exploring miscommunication in HRI. In *Procs. of the AISB Symposium on New Frontiers in Human-Robot Interaction*, pages 1–8.
- Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4):323–340, September.
- Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. 1998. An architecture for opportunistic text generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 28–37.
- Verena Rieser, Ivana Kruijff-Korbayová, and Oliver Lemon. 2005. A Corpus Collection and Annotation Framework for Learning Multimodal Clarification Strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.
- David Schlangen and Raquel Fernández. 2006. Beyond Repair Testing the Limits of the Conversational Repair System. In *7th 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*, pages 71–76.
- Dimitris Vogiatzis, Constantine D. Spyropoulos, Stasinios Konstantopoulos, Vangelis Karkaletsis, Zerrin Kasap, Colin Matheson, and Olivier Deroo. 2008. An affective robot guide to museums. In *Proceedings of the 4th International Workshop on Human-Computer Conversation, Bellagio, Italy*.
- M Walker, C Kamm, and D Litman. 2000. Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3&4):363–377.
- Teresa Zollo. 1999. A study of human dialogue strategies in the presence of speech recognition errors. In *Psychological Models of Communication in Collaborative Systems. Papers from the 1999 AAAI Fall Symposium (TR FS-99-03)*, pages 132–9. AAAI Press.