ABSTRACT
We present the evaluation of the most recent version of the Dutch ARISE train timetable information system. The original version of this spoken dialogue system has been adjusted according to the findings of two user tests [1,2]. The new version applies a mixture of implicit and explicit confirmation of information items, based on confidence measures. In addition, the negotiation part of the dialogue tells the user explicitly what he can ask. Furthermore, the exceptions handling was made very explicit. The new dialogue has been evaluated by 25 experts and by 200+ anonymous calls to the system. To be able to compare the two versions of the system, the same scenarios as in [1] were used. It was shown that the mixture of implicit and explicit confirmation results in shorter dialogues and in slightly higher dialogue success rates. Also, we observed a better performance in the negotiation part of the dialogue. However, the shortcomings of working with complicated scenarios are once again made clear.

Keywords: dialogue management, confidence measures, evaluation.

1 INTRODUCTION
The Dutch ARISE (Automatic Railway Information Systems for Europe) system is a spoken dialogue system for access to train timetable information of the Dutch Railways. The system was designed, developed and tested within the LE-3 ARISE project. It is closely related to an existing prototype system that is operational and available for customers (VIOS).

In [1] we reported on an evaluation of the Dutch ARISE system based on subjective and objective measurements of 68 users. In [2] the Dutch ARISE system was compared to the VIOS system. The results of both studies led to several major adjustments to the dialogue. The most important changes concerned the confirmation strategy and the negotiation part of the dialogue.

This paper presents the results of a user test and an expert evaluation that were carried out to test the new dialogue. First, in section 2 the architecture of the ARISE system and the changes to the dialogue manager are described. In addition, we present the set-up of the expert and user evaluations. In section 3, the results of the evaluations are presented. Finally, section 4 contains the discussion and conclusions.

2 METHOD

2.1 System architecture
The ARISE configuration consists of three components: (1) a Windows® NT workstation supporting the speech recognizer, the NLP-module and the dialogue manager, (2) a database server for the schedule information and (3) a telephony server. The speech recognition component consists of a speaker independent continuous speech recognizer, which uses phone based HMMs, unigram and bigram language models, and a 1044 word lexicon, including about 500 Dutch station names. The recognized sentences in the form of a word graph are passed to the NLP component which uses a unigram and bigram stochastic concept language model to determine the most likely sentence and to determine the meaning of this sentence. The dialogue manager controls the dialogue with the user and takes care that all slots for the database query are filled. All system prompts are generated by concatenating pre-recorded synthesized words and phrases.

2.2 Dialogue manager
As implicit confirmation has proven to be difficult for many people (they tend to get confused by the combination of confirmation request and a question for additional information), in previous versions of the ARISE system each information item was confirmed explicitly in separate turns. The user test described in [1] showed that subjects found explicit confirmation acceptable, if only because it was easier to correct errors. Explicit confirmation did not lengthen the dialogue duration (in seconds) compared to previous versions which used implicit confirmation throughout, mainly because very short prompts were used. However, it did increase the number of user turns and users found the dialogues more tedious. To reduce the number of turns and make the dialogue more natural, in the new version of the ARISE system a mixture of implicit and explicit confirmation is applied. The choice for implicit or explicit confirmation is based on the confidence that the information item has been correctly recognized. For this aim the algorithm discussed in [3] is used. If the system is very confident about the recognition result, the item is confirmed implicitly. The item is then stated before the next question is asked, see the two examples in (1).
If the user does not correct the system, the item is assumed to be confirmed and frozen. In case of a low confidence level, the item is confirmed explicitly in a separate question.

Furthermore, we implemented a rather explicit exceptions handling strategy throughout the dialogue. If the system does not understand the input, or if there was no input at all, the system helps the user by giving clear hints about the answering possibilities, e.g. ‘say the arrival station’. Previously, an extra question ‘what did you say?’ was asked first, but this often resulted in an exact repetition of the users utterance, which seldom solves the understanding problem.

Analyses of calls to the operator based service showed that most people ask for a connection for the present day and know the arrival time; therefore the system assumes the defaults ‘today’ and ‘arrival time’. We also changed the manner in which the travel advice is presented. In previous versions of the system, the advice gave the departure and arrival stations, final arrival and departure times and names and departure times at transfer stations. The new version of the ARISE system also provides information on arrival times at the transfer stations.

Once the travel advice has been presented, the user can navigate and negotiate when (s)he is not satisfied with the advice. The user can ask for an earlier/later connection, or a connection with less changes, for a faster connection, and for platform and direction information. The user can also ask for information about a return trip or another connection. Tests [1,2] showed that users easily get lost in this part of the dialogue, as they are not aware of the functionality of the system. Therefore this part of the dialogue was adjusted in a number of respects. Previously, after the presentation of the travel advice, the user was asked ‘is this the desired travel advice?’. As it turned out that the users’ answers to this question could not be interpreted unequivocally, this question has been changed to the more explicit ‘have you received sufficient information?’. In previous versions of the system, a negative answer to the first question evoked the open question: ‘which information do you want?’. In case the system did not receive relevant information, the user entered a menu in which all options were offered. Users had problems with answering the open question, which often led to very long user utterances, introducing a lot of recognition errors. Therefore, in the new dialogue, after a negative answer to the first question the user immediately enters a menu. The menu itself was made more rigid and straightforward. Also, the wording of the questions was changed to be more clear and unambiguous. Of course, the user can always take the initiative and ask for other information than (s)he is prompted for, in which case (s)he does not enter the menu. Furthermore, the dialogue has been made more ‘intelligent’: options that are not available are not offered at all. For example, the system no longer asks ‘do you want a connection with less changes’, when there is no such connection. Finally, an extra question was added, which is always asked: ‘do you want a repetition of the advice?’.

2.3 Scenarios and tasks

In order to be able to compare the results of the present test with those of the previous user test, the same scenarios were used. The scenarios were presented in a mixture of graphics and text, in order to avoid suggesting specific formulations, while at the same time eliciting specific user behavior.

The first scenario consisted of only one task, the other two scenarios were more difficult and consisted of three tasks each:

- **Scenario I**: a simple trip from A → B, following the default values ‘today’ and ‘arrival time’, cf. Figure 1.
- **Scenario II**: a trip from A → B, tomorrow, with arrival time two o’clock. The advice that was presented encouraged the user to ask for a later connection, as the arrival time of the ‘best’ connection was 59 minutes earlier than the designated arrival time. The user also had to ask for a return trip on the same day, with departure time eight o’clock in the evening.
- **Scenario III**: a trip from A → B in which both default values had to be changed. The scenario encouraged the user to ask for a connection with less changes (users were told that they had heavy suitcases and were presented a connection in which they had to change trains four times) and to ask for the exact changing times and for platform information.

2.4 Procedure

For the expert evaluation, 25 experts in the field of speech technology and dialogue systems were asked to complete the three scenarios. They were asked to write down the travel advices on an answer sheet, to ensure that they would listen carefully to the advice. They were also asked to fill in a questionnaire consisting of seven open questions concerning different aspects of the system.

Furthermore, more than 200 calls of more naive subjects, all (partners of) colleagues and students at Nijmegen University and KPN Research, were recorded. Log files of all dialogues were used to measure the objective performance of the system.

For each call to the system the following information was stored:

(1) a ‘From Amsterdam to Rotterdam. Today?’
  b ‘Today. At what time?’
The recognition performance of the ARISE system is described in terms of error rates that indicate the percentage of words, concepts or attributes that were recognized incorrectly (substituted, deleted or inserted). The word error rate (WER) denotes the number of incorrectly recognized words. The WER for this corpus is 25.1%. The number of out-of-vocabulary words in the corpus is 1.78%. Since not all words used in an utterance are equally important for a correct interpretation of the utterance, categories of words with similar meanings have been defined in a grammar, e.g. ‘station names’. These categories are called concepts. The concept error rate is 14.3%. Each meaningful concept has one or more attributes. The value of the attribute denotes the semantic content of the concept, for example the name of the station. The attribute error rate is 16.8%. Of all points in the dialogue where implicit confirmation could have been applied, it actually was applied in 56%. 6% of all implicitly confirmed items were in fact incorrectly recognized.

Dialogue success rate
Each of the three scenarios consisted of one or more tasks, which could be completed successfully (according to the description of the scenario), completed with wrong data or not completed at all. Table 1 shows the dialogue success rate per task. The dialogue success rate is the number of tasks that have been completed successfully as a percentage of the total number of dialogues. Dialogues that were completed with wrong data were not considered during the calculation of the dialogue success rate, as it is not clear whether the wrong data were due to system errors or the inattentiveness of the user.

3 RESULTS

3.1 Results of the objective evaluation
For the objective performance measures the dialogues of the 25 experts and the 200+ anonymous calls to the system were used.

Recognition performance
The recognition performance of the ARISE system is described in terms of error rates that indicate the percentage of words, concepts or attributes that were recognized incorrectly (substituted, deleted or inserted). The word error rate (WER) denotes the number of incorrectly recognized words. The WER for this corpus is 25.1%. The number of out-of-vocabulary words in the corpus is 1.78%. Since not all words used in an utterance are equally important for a correct interpretation of the utterance, categories of words with similar meanings have been defined in a grammar, e.g. ‘station names’. These categories are called concepts. The concept error rate is 14.3%. Each meaningful concept has one or more attributes. The value of the attribute denotes the semantic content of the concept, for example the name of the station. The attribute error rate is 16.8%. Of all points in the dialogue where implicit confirmation could have been applied, it actually was applied in 56%. 6% of all implicitly confirmed items were in fact incorrectly recognized.

 Dialogue success rate
Each of the three scenarios consisted of one or more tasks, which could be completed successfully (according to the description of the scenario), completed with wrong data or not completed at all. Table 1 shows the dialogue success rate per task. The dialogue success rate is the number of tasks that have been completed successfully as a percentage of the total number of dialogues. Dialogues that were completed with wrong data were not considered during the calculation of the dialogue success rate, as it is not clear whether the wrong data were due to system errors or the inattentiveness of the user.

Table 1 Dialogue success rate per task (between brackets the figures for the previous user test [1])

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Success</th>
<th>Wrong</th>
<th>Not compl.</th>
<th>Total</th>
<th>% Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Amsterdam Rotterdam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A → B</td>
<td>93</td>
<td>4</td>
<td>4</td>
<td>101</td>
<td>96% (93%)</td>
</tr>
<tr>
<td>II. Zwolle Hindeloopen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A → B</td>
<td>101</td>
<td>2</td>
<td>9</td>
<td>112</td>
<td>92% (81%)</td>
</tr>
<tr>
<td>later</td>
<td>66</td>
<td>0</td>
<td>5</td>
<td>71</td>
<td>93% (88%)</td>
</tr>
<tr>
<td>B → A</td>
<td>69</td>
<td>10</td>
<td>16</td>
<td>95</td>
<td>81% (43%)</td>
</tr>
<tr>
<td>III. Alkmaar Maastricht</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A → B</td>
<td>81</td>
<td>2</td>
<td>6</td>
<td>89</td>
<td>93% (85%)</td>
</tr>
<tr>
<td>Less changes</td>
<td>62</td>
<td>0</td>
<td>6</td>
<td>68</td>
<td>91% (85%)</td>
</tr>
<tr>
<td>Platform info</td>
<td>58</td>
<td>1</td>
<td>8</td>
<td>67</td>
<td>88% (69%)</td>
</tr>
<tr>
<td>Total</td>
<td>530</td>
<td>19</td>
<td>54</td>
<td>603</td>
<td>91% (78%)</td>
</tr>
</tbody>
</table>

Table I shows that the overall success rates are higher than in the previous test. This may be partly due to the fact that subjects were less naive than in the previous test. In the present test the subjects were experts, colleagues and students, whereas in the previous test the subjects were customers of the operator-based service. The first task of scenario I was most successful. In this scenario no defaults had to be changed, whereas in scenarios II and III one or two defaults had to be changed. The high success rate for the third task of scenario II, compared to the previous test, can be explained by the fact that a system bug that caused the low success rate in the previous test [1] has been solved. That not all people completed the second and third task of scenarios II and III was also observed in the previous test [1]. In this test we observed ten subjects at the lab and we have noticed that some users did not understand that they were supposed to carry out this task due to the manner in which the scenario was presented. Other users simply did not know how to ask for the required information.

Number of turns
For the successful dialogues, the minimum, modal and maximum number of turns was calculated from the first user utterance until the presentation of the travel advice. In a mixed initiative dialogue, the minimum number of turns strongly depends on the amount of initiative the user takes. In the case of a cooperative user who provides all necessary information in the first turn, the minimum number of turns for the first part of a dialogue within the ARISE system is three (provided that everything has been correctly understood by the system and with a high confidence). In the more likely case where the user answers the questions asked by the system exactly and provides no extra information, the theoretical minimum number of turns is four for scenario I, five and six for scenarios II and III respectively.
Table 2 Number of turns in successful dialogues (between brackets the figures for the previous user test [1])

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Min</th>
<th>Modus</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Amsterdam Rotterdam</td>
<td>3</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>II. Zwolle Hindebloopen</td>
<td>3</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>III. Alkmaar Maastricht</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2 shows that for each scenario the minimum number of turns observed is equal to the theoretical minimum number of turns. The first task was completed in three turns in almost 12% of all successful dialogues. The modal number of turns turned out to be four for each scenario. For scenarios II and III the modal number of turns is lower than we expected. This is explained by the fact that people do use mixed initiative in case they have to change a default value. Instead of just answering ‘no’ to the question ‘today?’, they answer ‘no, tomorrow’. If the correction is confirmed implicitly, no extra turn is necessary.

Table 2 also shows that, thanks to implicit confirmation, the number of turns decreased considerably compared to the previous test. That the difference is larger for scenarios II and III is explained by the fact that changing the defaults need not cost any extra turns any more. Despite the fact that implicit confirmation was only applied in 56% of the nodes where it could have been applied (see section 3.1, Recognition performance), its influence is large. This can be explained by the fact that the successful dialogues contain relatively more implicit confirmations than the unsuccessful dialogues.

The number of turns in the negotiation part of the dialogue for scenario II and III also decreased, even though an extra question was asked.

3.2 Results of the questionnaire

Although most experts were familiar with the original ARISE system, few people noticed the changes in the confirmation strategy. Those who did notice the difference found it acceptable. The more explicit exceptions handling mechanism was much appreciated. It is perceived as straightforward and preventing long dialogues.

The second part of the dialogue remains difficult. The mental model people have of the interaction with the machine does not match reality. It is very difficult, if not impossible, to make the user aware of the possibilities and limitations of the system. The menu-structured dialogue helps to make the system more transparent, but it makes the dialogue too tedious for those people who received the desired travel advice at once and want to leave the service. People seemed to be reluctant to simply hang up the phone in the second part of the dialogue, before the system has proceeded to its farewell message.

Finally, the fact that the dialogue was mixed initiative was often not noticed, and the possibility to provide the system with more information than the user is prompted for was rarely used. This is probably because the questions that are asked by the system are fairly directive. Besides, often not all extra information is correctly recognized, which can be interpreted by the user as if it is not possible to provide more information than the system asked for.

4 DISCUSSION & CONCLUSION

We adapted the dialogue management of the Dutch ARISE system according to the results of the evaluations of the previous version of the ARISE system. The present evaluation shows that the users consider the combination of implicit and explicit confirmation, the changes to the travel advice and the more explicit exceptions handling, to be improvements. Also, the system performance improved because of the changes made to the system. The changes resulted in a higher dialogue success rate and a lower number of turns in all scenarios. However, the algorithm used to calculate the confidence measures needs improvement, in order to decrease the number of incorrect implicit confirmations. A number of difficulties remain unsolved. We attempted to make the negotiation part of the dialogue more transparent and straightforward. Yet, this part still turned out to be difficult for many people. Although users were offered more guidance to discover what the possibilities were, many people did not succeed in completing the more complex scenarios. Another group of people considered the second part of the dialogue to be too tedious, due to the more directive questions.

The evaluation also showed once again that the use of complicated scenarios to evaluate a dialogue system causes problems. Although using scenarios makes the situation controllable, it is very difficult to make the user perform more complex tasks. Finally, the poor recognition performance remains a bottleneck in a spoken dialogue system. Since the mixed initiative possibilities of the system are not much used, a system driven dialogue might be a solution for the recognition problems. The more so because this offers the possibility to adapt the lexicon, the language models and the acoustic models that are used during recognition, to the question that is asked, which has proven to result in a higher word accuracy.

5 REFERENCES