D6. 4: Final evaluation of CLASSiC TownInfo and Appointment Scheduling systems

Article · May 2011

CITATIONS
15

READS
56

11 authors, including:

Helen Hastie
Heriot-Watt University
105 PUBLICATIONS **858** CITATIONS
SEE PROFILE

Filip Jurcicek
Charles University in Prague
55 PUBLICATIONS **439** CITATIONS
SEE PROFILE

Oliver Joseph Lemon
Heriot-Watt University
323 PUBLICATIONS **3,678** CITATIONS
SEE PROFILE

Steve Young
University of Cambridge
310 PUBLICATIONS **14,308** CITATIONS
SEE PROFILE

CLASSiC

D6.4: Final evaluation of CLASSiC TownInfo and Appointment Scheduling systems

Romain Laroche, Ghislain Putois, Philippe Bretier,
Martin Aranguren, Julia Velkovska, Helen Hastie, Simon Keizer,
Kai Yu, Filip Jurcicek, Oliver Lemon, Steve Young

Distribution: Public

CLASSiC
Computational Learning in Adaptive Systems for Spoken Conversation
216594 Deliverable 6.4

April 2011

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref. no.	216594
Project acronym	CLASSiC
Project full title	Computational Learning in Adaptive Systems for Spoken Conversation
Instrument	STREP
Thematic Priority	Cognitive Systems, Interaction, and Robotics
Start date / duration	01 March 2008 / 36 Months

Security	Public
Contractual date of delivery	M36 = February 2011
Actual date of delivery	April 2011
Deliverable number	6.4
Deliverable title	D6.4: Final evaluation of CLASSiC TownInfo and Appointment Scheduling systems
Type	Report
Status & version	Draft 1.0
Number of pages	83 (excluding front matter)
Contributing WP	6
WP/Task responsible	FT, UCAM
Other contributors	HWU, SUPELEC
Author(s)	Romain Laroche, Ghislain Putois, Philippe Bretier, Martin Aranguren, Julia Velkovska, Helen Hastie, Simon Keizer, Kai Yu, Filip Jurcicek, Oliver Lemon, Steve Young
EC Project Officer	Philippe Gelin
Keywords	Spoken dialogue systems, system evaluation

The partners in CLASSiC are:

Heriot-Watt University	HWU
University of Cambridge	UCAM
University of Geneva	GENE
Ecole Superieure d'Electricite	SUPELEC
France Telecom/ Orange Labs	FT
University of Edinburgh HCRC	EDIN

For copies of reports, updates on project activities and other CLASSiC-related information, contact:

The CLASSiC Project Co-ordinator:
Dr. Oliver Lemon
School of Mathematical and Computer Sciences (MACS)
Heriot-Watt University
Edinburgh
EH14 4AS
United Kingdom
O.Lemon@hw.ac.uk
Phone +44 (131) 451 3782 - Fax +44 (0)131 451 3327

Copies of reports and other material can also be accessed via the project's administration homepage, `http://www.classic-project.org`

Contents

Executive summary

This document is a report on the final evaluations of the CLASSiC TownInfo and Appointment Scheduling systems. It describes the setup and results of the experiments involving real users calling 4 different systems to perform different tasks and give ratings to each dialogue. For both TownInfo and Appointment Scheduling (AS) domains, one of the evaluated systems incorporated several components from different sites within the consortium. For more details about these integrated systems, see D5.2.2 for the CLASSiC TownInfo systems, and D5.4 for the CLASSiC Appointment Scheduling systems.

For the TownInfo evaluations a total of 2046 dialogues were collected. For the AS systems, System 2 collected a total of 628 dialogues, while Systems 3 and 4 collected 740 and 709 dialogues for evaluation respectively, for a total of 2077 AS dialogues.

The main contrasts explored in the TownInfo evaluations were the effects of processing N-best lists as input to the dialogue system (using POMDP techniques) as opposed to using only 1-best ASR input, and the effects of using the trained NLG components.

The AS evaluation explores the differences between several systems:

- the 'academic' system, with and without a trained NLG component (System 2)
- the FT commercial system that was adapted to the experimental set-up (System 3)
- the FT lab system that is an evolution of the FT commercial system using questions that do not constrain the user in a predefined behaviour. This system embeds also uncertainty management. (System 4)

Part I of the report concerns the TownInfo system (System 1) and Part II concerns the Appointment Scheduling systems (Systems 2, 3, and 4) This report also presents the sociological evaluation of the Appointment Scheduling systems carried out by France Telecom / Orange Labs (Part II, Chapter 5).

Results from the TownInfo trial were mixed. Four main measures were applied: subjective success rate (PercSucc), objective partial completion based on the assigned goals (ObjSucc-AG-PC), objective full completion based on the assigned goals (ObjSuc-AG-FC), and objective full completion based on the inferred goals (ObjSucc-IG). Partial completion requires only that subjects found an appropriate venue whereas full completion required that they obtained all of the required ancillary information such as phone number and address. The inferred goals (IG) measure attempted to match the system's responses to what the user actually asked for, rather than the assigned goals.

On partial completion, the CLASSiC system with the specialised NLG component was significantly better than the other systems. On the remaining measures, the systems were broadly similar in performance. A striking feature of all the results was that the objective measures were all much lower than the subjective success rates (PercSucc). This is thought to be mostly because users were often unaware that the venue offered did not actually satisfy their goals or that they had failed to ask for certain required information. This illustrates one of the major shortcomings of this type of trial.

One surprising feature of the TownInfo trial results was that in contrast to the simulation results, in several cases the N-best system did not perform better than the 1-best system. Evaluation of semantic accuracy indicated that there was additional information in the N-best lists from the recogniser but clearly the dialogue manager failed to exploit it. The most likely reason for this is that the error model used in the user simulator is a poor match to the actual pattern of errors incurred in the real data. This reinforces the need to move away from training on simulators and instead training on real user data.

A major performance issue with the TownInfo trial arose from the lack of appropriate training data. This resulted in a system in the main Feb'11 trial with a word error rate ranging from 53% to 56%. However, even with these very high WERs, perceived success rates of 60% to 65% were achieved in the Feb'11 trial. This shows that the systems were fairly resilient even when operating in extremely hostile conditions. Following the trial, the data collected was used to retrain the recogniser with the result that the error rate was halved to a WER of 26%. A further trial was then conducted after the project officially ended, and the perceived success rate increased to 88%, showing the impact of the poorly trained recognition models.

Three different systems for Appointment Scheduling (AS) were also evaluated (Systems 2, 3, and 4), using over 2000 dialogues. System 3 is a variant of the deployed France Telecom 1013+ service, and System 4 is a more advanced laboratory version of this system. System 2 was built using the statistical components developed by the academic partners in the project. Although comparing Systems 2, 3 and 4 directly is not possible due to the different speech recognition components used, we can draw some general conclusions about the comparative performance of the different systems.

While commercial systems are typically deployed only after many iterations of user testing. In this case, both System 2 and System 4 were trialled following minimal testing, and achieved comparable performance to System 3 (all performing at around 80% task completion). System 3 was already the result of on-line optimisation, which resulted in a 10% task completion increase. This means that Systems' 2 and 4 performances already exceed classical handcrafted performance. In addition, these systems were developed rapidly using the methods and tools developed during the CLASSiC project.

Regarding the trained NLG component, the version of System 2 which included the trained component for Temporal Referring Expression generation showed a statistically significant improvement in Perceived Task Success (+23.7%) and a reduction in call time of 15.7% (to appear, [20]).

The issue of how much freedom it is beneficial to give the user (i.e. user- or system-initiative) is also explored in detail, in section 4.4.

Chapter 5 also presents further detailed qualitative analysis of the AS dialogues using methods from Conversation Analysis, for example examining types of errors and interactional misalignment phenomena between the user and the system. This leads to suggestions of strategies for error recovery.

Taken together, this set of results shows that the statistical learning methods and tools developed in the CLASSiC project provide a promising foundation for future research and development into robust and adaptive spoken dialogue systems.

Chapter 1

Introduction

This document is a report on the final evaluations of the CLASSiC TownInfo and Appointment Scheduling systems. It describes the setup and results of the experiments involving real users calling different systems to perform different tasks and give ratings to each dialogue.

Part I of the report concerns the TownInfo system (System 1) and Part II concerns the Appointment Scheduling systems (Systems 2, 3, and 4) This report also presents the sociological evaluation (using Conversation Analysis) of the Appointment Scheduling systems carried out by France Telecom / Orange Labs (Part II, Chapter 5).

For the TownInfo systems, the actual domain was switched from an imaginary town ('Jasonville') to real locations in Cambridge, and VoIP technology was used during evaluation, resulting in the 'CamInfo' system. Subjects were asked to find a place to eat in Cambridge, following a scenario given to them. This change added a lot more realism and also achieved greater efficiency of carrying out large scale dialogue system evaluation.

For the Appointment Scheduling systems, the subjects were asked to book an appointment on one of the free slots in a user calendar given to them. Systems built by France Telecom and the academic CLASSiC team were evaluated on the same tasks.

For both the TownInfo and Appointment Scheduling domains, one of the evaluated systems used components contributed by different sites within the consortium. For more details about these integrated systems, see deliverable D5.2.2 for the CLASSiC TownInfo system, and deliverable D5.4 for the CLASSiC Appointment Scheduling system.

For the TownInfo evaluations (Part I) a total of 2046 dialogues were collected. For the AS systems, System 2 collected a total of 628 dialogues, while Systems 3 and 4 collected 1449 dialogues for evaluation.

The main contrasts explored in the TownInfo evaluations were the effects of processing N-best lists as input to the dialogue system (using POMDP techniques) as opposed to using only 1-best ASR input, and the effects of using the trained NLG components.

The AS evaluation (Part II) explores the differences between the 4 different Appointment Scheduling systems developed in the project:

- the 'academic' system, with and without a trained NLG component (System 2)
- the FT commercial system that was adapted to the experimental set-up (System 3)
- the FT lab system that is an evolution of the FT commercial system using questions that do not con-

strain the user in a predefined behaviour. This system also uses uncertainty management. (System 4)

We note that System 2 cannot be directly compared with System 3 and System 4, since different speech recognisers were used. However, we can draw some general lessons and conclusions about the performance of the CLASSiC systems and the methods used to develop them (see Chapter 6).

The issue of how much freedom it is beneficial to give the user (i.e. user- or system-initiative) is also explored in detail, in section 4.4.

In Chapter 5 we also present further detailed qualitative analysis of the AS dialogues using methods from Conversation Analysis, for example examining types of errors and interactional misalignment phenomena between the user and the system. This leads to suggestions of strategies for error recovery.

Part I

TownInfo evaluations

Chapter 2

Final TownInfo evaluations

2.1 Introduction

In deliverable D6.3, the setup and results of the initial evaluation of the CLASSiC TownInfo system were presented. The experiments at that time involved subjects who were recruited to come into the lab and talk to different systems using a desktop computer and headset. The experience gained from this preliminary evaluation indicated a number of significant problems. Firstly, the whole process was labour intensive and the number of subjects that could be managed was therefore limited. Furthermore, the task and conditions of the trial were artificial and did not represent real world operating conditions. The imaginary town 'Jasonville' was very small and venues could be located rapidly with just one or two search constraints, the use of close talking microphones was unrepresentative and the user behaviour was as a result probably atypical.

With the aim of obtaining more realistic conditions, the TownInfo system was subsequently ported from the relatively small tourist information domain involving a fictitious town to the much larger, and real-world, Cambridge tourist information domain. In addition, the system was integrated with a VoIP server to allow phone connections from anywhere in the world. In this way, subjects were no longer required to come to the lab to do the experiment, and the trial setup was more like real usage.

However, there were a number of issues that arose from using this new more realistic test environment. Firstly, the original artificial TownInfo 'Jasonville' domain had been subjected to a number of evaluations and these had provided good quality data for training the language model used by the recogniser. No such data was available for the Cambridge domain so instead artificial data was constructed by mapping the Jasonville data. Secondly, there was no acoustic training data which matched the VoIP audio channel and hence the Jasonville training data had to be reused. Thirdly, the larger catchment area for subjects and the need to increase the number of test dialogues made it more difficult to control the origin and accent of the users. Thus, the speaker variability in this new setup was much wider than had been captured by our existing training data. As will be seen in the results below, the net outcome was a dialogue system which was required to operate with very high error rates.

2.2 Experimental setup

Subjects were recruited using mail-shots and web-based advertising amongst people from Cambridge as well as Edinburgh, mostly students. From the resulting pool of subjects, people were gradually invited to start doing tasks in their own time within a given trial period of around two weeks. After the trial period, they were paid per completed task, with a required minimum of 15 tasks, and a maximum of 40 tasks. In total, 1124 evaluation dialogues were collected in this way.

When invited, the subjects were pointed to a website with detailed instructions and for each task, a phone number to call (corresponding to one of the three systems evaluated) and the scenario to follow. All scenarios described a place to eat in Cambridge with some additional constraints, for example: "You want to find a moderately priced restaurant and it should be in the Fen Ditton area. You want to know the address, phone number, and type of food.". After the dialogue, the subjects were asked to fill in a short questionnaire:

> Q1. Did you find all the information you were looking for? [Yes / No]
>
> Please state your attitude towards the following statements:
>
> Q2. The system understood me well. [1 – 6]
> Q3. The phrasing of the system's responses was good. [1 – 6]
> Q4. The system's voice was of good quality. [1 – 6]
>
> | | |
> |---|---|
> | 1: strongly disagree | 4: slightly agree |
> | 2: disagree | 5: agree |
> | 3: slightly disagree | 6: strongly agree |

In addition to the setup described above, an alternative method of recruiting and managing subjects was used, using Amazon Mechanical Turk. In this setup, tasks are published as so-called HITs (Human Intelligence Tasks) on a web-server and registered workers can complete them. This setup resulted in 922 collected dialogues, so in total, 2046 dialogues were collected for the final TownInfo evaluation.

2.2.1 Evaluated systems

All three systems included in the evaluation shared the same speech recogniser and semantic parser (SLU), and dialogue manager (DM), all developed at Cambridge University. For the speech synthesis (TTS), the France Telecom Baratinoo synthesiser was used, again in all three systems.

One of the systems is the CLASSiC integrated System 1, in which the NLG module has a recommendation component, developed by Edinburgh University. This component is called only when the dialogue manager decides to make a venue recommendation. It chooses one of several possible 'recommend' actions using a trained generation policy and returns the generated utterances.

Table 2.2.1 gives an overview of the three systems.

The other two systems differ only in their dialogue management policies[1]. During the experiment, one system was given the full N-best list of user act hypotheses, resulting from the 10-best ASR hypotheses, while the other system only had the 1-best semantic hypothesis at its disposal. To make a fair comparison, the policies for these two systems were trained on simulated data with corresponding noise-conditions.

[1] As in the initial TownInfo evaluation, the Hidden Information State dialogue manager [1] was used for all three systems.

System	SLU	DM	NLG	TTS
N-Best-UCAM	UCAM	UCAM	UCAM	FT
1-Best-UCAM	UCAM	UCAM	UCAM	FT
N-Best-CLASSiC	UCAM	UCAM	UEDIN/HW	FT

Table 2.2.1: Overview of evaluated systems.

One policy was trained using a semantic level error model producing N-best lists of semantic hypotheses, and the other using only 1-best lists of simulated user act hypotheses. Both policies were first evaluated on simulated data. The results in Figure 2.2.1 show that both POMDP policies profit from alternative hypotheses (see eval3best vs eval1best performance). Second, the 3-best policy outperforms the 1-best policy on 3-best simulated data (see pol3best-eval3best vs pol1best-eval3best performance) and vice versa, the 1-best policy outperforms the 3-best policy on 1-best simulated data (see pol1best-eval1best vs pol3best-eval1best performance). Finally, looking forward to the real user evaluation, the 3-best policy operating with 3-best semantic input outperforms the 1-best policy operating with 1-best semantic input (see pol3best-eval3best vs pol1best-eval1best performance).

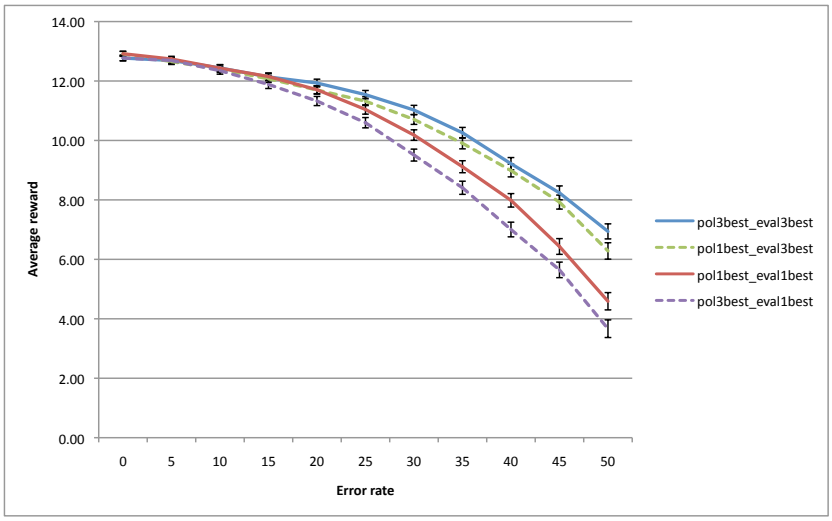

Figure 2.2.1: Evaluation on simulated data of the 3-best and 1-best HIS policies used in the Nov'10 evaluation.

The first evaluation period in November 2010 was suspended before the targeted number of dialogues was collected. The intermediate results suggested that there might be a problem with the systems and therefore an analysis of the data was necessary. Although no significant problems were discovered, the analysis suggested that the actual error rates being encountered were much higher than had been assumed when the policy was trained. Furthermore, the pattern of errors appeared to be different. This analysis resulted in the decision to retrain the policies using a modified error model for generating N-best lists of semantic hypotheses from simulated user acts. This new error model also enabled us to tune it to better match corpus data in terms of oracle rate and top accuracy. The results of the evaluation of the new policies on simulated data are given in Figure 2.2.2. In this case, the relative performance of the policies are similar, but only at higher error rates.

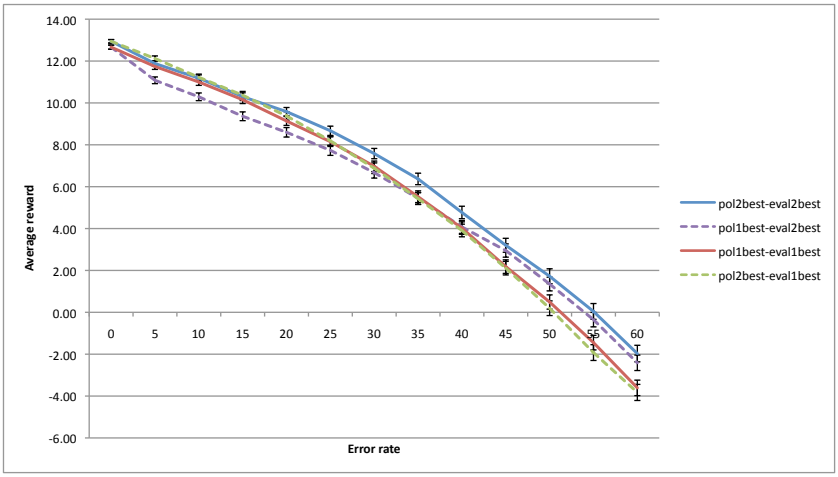

Figure 2.2.2: Evaluation on simulated data of the 2-best and 1-best HIS policies used in the Feb'11 evaluation.

2.2.2 Construction of the speech recogniser

As noted in the introduction, a major challenge in building the new VoIP based spoken dialogue system for the much larger Cambridge tourist information domain was the construction of the speech recogniser due to the lack of acoustic and language model training data.

The existing acoustic model for the TownInfo system (CLASSiC System 1) was trained on about 39 hours of wide-band audio from British native speakers. This could not be directly used for the Cambridge Information System due to the channel mismatch caused by the use of VoIP. To deal with this, the original wide-band models were converted to narrow-band models using single-pass retraining. At the same time, the cepstral mean and HLDA transform were also re-estimated. This provided the initial CamInfo VoIP acoustic models. Following that, about 1 hour of VoIP acoustic data was collected from lab users at Cambridge University. A new acoustic model was then trained with the 1 hour VoIP data using MAP adaptation. It is worth noting that the training data were mostly from native English speakers (except for a small portion from non-native speakers in the 1 hour of collected VoIP data). Thus, the acoustic models used for the VoIP system were not trained on any in-domain data and were not well-matched to the target domain where we subsequently encountered significant channel variability and a very high proportion of non-native speakers. However, given the time constraints and the significant cost involved in collecting substantial amounts of real in-domain data, there was no viable alternative. The outcome was that the mismatched acoustic models used in the CamInfo evaluation lead to a significant performance degradation during the evaluation.

Building a suitable language model for the Cambridge Information domain brought a similar lack-of-data problem since there were no transcriptions available for dialogues in this domain. Instead, a language model was bootstrapped from pseudo-data. First, about 600 in-domain text sentences were manually generated. Then the content words in these sentences were automatically replaced by similar content words from the Cambridge Information database. This automatically generated corpus had about 100K words. A tri-gram language model was trained on this small corpus and interpolated with the previous TownInfo domain tri-gram language model. The final interpolated tri-gram was then pruned and used for automatic speech recognition. Although it was ensured that all named entities in the Cambridge Tourist

Information database were in the language model, the language model itself was still a relatively poor language model because the in-domain training data was artificial and insufficient. Again, this severely affected the recognition performance of the trialled systems.

2.3 Results

In total, 2046 dialogues were collected: 709 dialogues in November 2010, and 1337 dialogues in February 2011. The most important metrics used in the real user evaluation are based on task completion/success. The subjective/perceived success rate is obtained from first question of the questionnaire: "Did you get all the information you were looking for?" (Q1, see Section 2.2). For computing objective success rates we currently have two alternative methods. The first method relies on the tasks given to the user for each dialogue (the 'assigned' goal: AG) and on the system dialogue acts. If the system has offered a venue matching the predefined constraints, the task is said to be 'partially completed' (PC); if the system has additionally provided all required information about the offered venue (e.g., the phone number and address), it is said to be 'fully completed' (FC). The problem with this method is that it requires the subjects to follow the scenarios closely. If they forget to mention a constraint to the system or forget to ask for some of the required information, the dialogue will be considered unsuccessful, although the system cannot really be blamed for it. For this reason, an alternative method for computing task success was developed, which does not rely on the predefined task, but on what the user actually asked for. Using a semantic transcription of the user utterances, the goal(s) are inferred from the system and user dialogue acts using fine-tuned heuristics, and task success is then based on these inferred goals (IG). For this evaluation, the semantic transcriptions were obtained by running the semantic parser on the audio transcriptions. The audio transcriptions themselves were obtained using transcribers recruited via the Amazon Mechanical Turk crowd-sourcing platform. Such semantic transcriptions were considered to be sufficiently reliable, because a study on a similar dialogue corpus collected previously showed that the performance of the semantic parser used was very similar to the quality of human annotations. Table 2.3.1 summarises the metrics used.

Metric	Description
PercSucc	Perceived/subjective Success rate
ObjSucc-AG-PC	Assigned Goal based Partial task Completion
ObjSucc-AG-FC	Assigned Goal based Full task Completion
ObjSucc-IG-FC	Full task Completion based on Inferred Goal

Table 2.3.1: Metrics used for measuring task success rates.

2.3.1 Effect of ASR

As indicated in Section 2.2.2, due to the lack of acoustic and language model training data, the speech recogniser used in the VoIP-based CamInfo system was far from optimal. The overall ASR performance for the evaluation is shown in Table 2.3.2.

From the table, it is clear that ASR performance was poor; indeed, for most commercial systems this level of performance would be regarded as unusable.

Evaluation	Subjects	WER (%)
Nov. 2010	Cambridge	44.4
Feb. 2011	MTurk	53.6
Feb. 2011	Cambridge	56.5

Table 2.3.2: Word Error Rates (WER) of Nov. 2010 and Feb. 2011 evaluations.

In fact, the lack of training data was not the only problem. During the evaluations, we discovered a number of factors which further contributed to poor performance. Firstly, a significant number of callers (8%) used Skype instead of landlines or mobile phones to call the system. Skype users typically use the far-field microphone in their laptop which then requires echo cancelling algorithms to be used. All of this distorts the audio signal. Secondly, in the Cambridge evaluation, a large portion of the subjects were non-native speakers due to difficulties with recruiting only native English speakers with British accents. However, in the Amazon Mechanical Turk (MTurk) evaluation subjects were English speakers with North American accents which mismatched the acoustic model trained on speakers with British accent. Furthermore, not only were most callers not native speakers, the speaking style was significantly more casual than had previously been encountered with lab-based users. There were also many non-speech events, such as coughs, sniffs, laughs, etc., which broke the silence detection and led to further recognition errors. These all contributed to poor automatic speech recognition (ASR) performance for the evaluations.

To demonstrate the impact of the lack of training data, following the evaluation, we cut out a small amount of the evaluation data to provide a held-out test set and then folded the rest into the acoustic training data and retrained the acoustic and the language model after obtaining the required orthographic transcriptions from Amazon MTurk. The total amount of new in-domain audio data made available for training was about 15 hours. Together with the original 1 hour of VoIP data, MAP training was performed to build a new set of acoustic models. The acoustic transcriptions made available for language model training consisted of 88K words. A separate tri-gram language model was trained from this corpus and then interpolated with the original Cambridge Information tri-gram model. The held-out evaluation data set comprised about 1.3 hours (150 dialogues) selected from the Feb. 2011 evaluation. The performance of using the new models tested on the held-out data set is shown in table 2.3.3:

System	AM	LM	
baseline	—	—	57.2
New Model	—	new	43.8
	new	new	26.5

Table 2.3.3: WER (%) comparison on held-out test set

It can be observed that folding in in-domain acoustic and language model training data halved the word error rate on the Feb 11 evaluation data to about 26% WER. This is much closer to the level of performance which is generally considered to be usable for a spoken dialogue system (10% to 30% is the normal target range for commercial systems).

To verify the quality of the improved ASR, an extra additional trial was conducted after the project finished. The trial used the N-Best-UCAM system with the new acoustic and language models with subjects

recruited using MTurk ie. mostly americans. After 274 dialogues had been collected the perceived success rate was 88% compared to the 64% obtained in the Feb'11 trial. This difference is statistically significant at the 95% confidence level when using a two-tailed z-test. The objective results will not be available until the dialogues have been transcribed, but given these subjective scores, it is reasonable to expect that all the success rate measures would increase similarly.

2.3.2 Evaluation 1: November 2010

The results in terms of success rates for both November 2010 and February 2011 evaluations are given in Table 2.3.4. Regarding the Nov'10 trial results, the 1-Best system outperformed the two N-Best systems in terms of the perceived success rate (PercSucc), in contrast to our expectation. The objective success rates based on full completion of assigned tasks (ObjSucc-AG-FC) confirm this result, though the absolute scores are much lower. These lower scores are partly caused by subjects not strictly following the scenarios given to them, resulting in overly pessimistic scores. The partial completion success rates (ObjSucc-AG-PC) are more similar to the perceived success rates, so typically subjects forgot to ask for certain information about an offered venue that was required in the scenario. The partial completion score for the N-Best-CLASSiC system (i.e. including the trained NLG component) was significantly higher than the N-Best-UCAM system (p=0.02, z-test), suggesting that the more elaborate venue offers from the EDIN/HWU-NLG helped the user find the venue they were looking for more easily. In terms of the completion rates based on inferred user goals (IG), the scores for the three systems are very similar. Again, the absolute scores are much lower than the subjective ratings (PercSucc) , suggesting a further bias in the user judgment, apart from the deviations from the scenario. The results from the other questions

Trial	System	NumDials	PercSucc	ObjSucc-AG-PC	ObjSucc-AG-FC	NumTasks	ObjSucc-IG
Nov'10	N-Best-UCAM	238	79.41 (±5.14)	78.57 (±5.21)$_1$	60.50 (±6.21)	250	65.60 (±5.89)
	1-Best-UCAM	212	86.32 (±4.63)	81.60 (±5.22)	70.28 (±6.15)	227	65.20 (±6.20)
	N-Best-CLASSiC	259	78.76 (±4.98)	**86.10** (±**4.21**)$_{\mathbf{1}}$	62.16 (±5.91)	278	64.03 (±5.64)
Feb'11	N-Best-UCAM	199	65.33 (±6.61)	73.37 (±6.14)	46.73 (±6.93)	214	43.93 (±6.65)
	1-Best-UCAM	111	62.16 (±9.02)	68.47 (±8.64)	38.74 (±9.06)	125	38.40 (±8.53)$_3$
	N-Best-CLASSiC	105	60.00 (±9.37)	77.23 (±8.02)	49.50 (±9.56)	118	**50.85** (±**9.02**)$_{\mathbf{3}}$
Feb'11 Mturk	N-Best-UCAM	402	64.18 (±4.69)	51.00 (±4.89)$_2$	28.86 (±4.43)$_4$	428	44.63 (±4.71)
	1-Best-UCAM	390	67.44 (±4.65)	58.21 (±4.90)	36.67 (±4.78)	425	50.35 (±4.75)
	N-Best-CLASSiC	130	56.15 (±8.53)	**60.77** (±**8.39**)$_{\mathbf{2}}$	**37.69** (±**8.33**)$_{\mathbf{4}}$	151	47.68 (±7.97)

Table 2.3.4: Subjective and objective success rates. NumDials is the number of dialogues; NumTasks is the number of tasks corresponding to inferred goals, where multiple goals can be inferred from a single dialogue. For the given success rates, 95% confidence intervals are indicated between brackets; see Table 2.3.1 for an overview of the various task success metrics. (1, 2, 3 = statistically significant differences at $p < 0.05$)

in the questionnaire, described in Section 2.2, are given in Table 2.3.5. For all those questions, all three systems have similar performance, though there seems to be a slight correlation with the perceived success rate.

2.3.3 Evaluation 2: February 2011

After retraining the dialogue manager policies with the new error model, the experiment was resumed. In order to do this, new subjects had to be recruited. This time, in addition to local recruitment, an alternative

Trial	System	NumDials	PercSucc (Q1)	PercUnd (Q2)	PercPhr (Q3)	PercVoi (Q4)
Nov'10	N-Best-UCAM	238	79.41	4.40	4.66	4.71
	1-Best-UCAM	212	86.32	4.46	4.66	4.78
	N-Best-CLASSiC	259	78.76	3.99	4.26	4.49
Feb'11	N-Best-UCAM	199	65.33	3.69	3.94	4.23
	1-Best-UCAM	111	62.16	3.36	3.85	4.15
	N-Best-CLASSiC	105	60.00	3.44	3.70	3.91
Feb'11 Mturk	N-Best-UCAM	402	64.18	3.92	4.16	3.82
	1-Best-UCAM	390	67.44	3.94	4.22	3.89
	N-Best-CLASSiC	130	56.15	3.87	4.30	3.85

Table 2.3.5: Evaluation scores for the perceived performance of understanding/SLU (PercUnd), phrasing/NLG (PercPhr), and voice/TTS (PercVoi).

method of recruiting and managing subjects was used. In this setup, crowd-sourcing technology is used for the recruitment, instruction and payment of subjects. Using Amazon Mechanical Turk (MTurk), tasks are published as so-called HITs (Human Intelligence Tasks) on a web-server and registered workers can complete them. The main difference in these setups is in the subject populations: the Cambridge based setup involved both British native speakers and non-native speakers, whereas the Mechanical Turk setup mostly attracted American native speakers.

The results in terms of the various types of success rate are shown in Table 2.3.4. As in the November evaluation, the objective scores based on assigned tasks are much lower than the subjective ratings, and this is particularly the case in the MTurk experiment. The MTurk subjects seem to be less likely to follow the task than the Cambridge recruited subjects. Furthermore, the objective scores do not confirm the subjective ratings this time. It should also be noted that the subjective scores for the MTurk evaluation might be particularly biased because of the relatively low average number of tasks each user completed. The statistics in Table 2.3.6 shows that on average, Cambridge recruited subjects, who were required to do a minimum number of tasks to get paid, completed significantly more tasks than MTurk subjects, who were not restricted in this way.

Trial	NumUsers	NumDials	AvgNumDialsPerUser
Nov'10	28	709	25.32
Feb'11	19	415	21.84
Feb'11 MTurk	113	922	8.16

Table 2.3.6: Number of different users, number of dialogues, and average number of dialogues per user, for each of the three trials.

As in the Nov'11 evaluation, in the Feb'11 MTurk evaluation, the N-Best-CLASSiC system (i.e. with trained NLG) significantly outperforms the N-Best-UCAM system on partial completion.

In the Cambridge based Feb'11 evaluation, the inferred goal based success rates for the N-Best systems are better than those of the 1-Best system. The success rate for the N-Best-CLASSiC system is significantly higher than that for the 1-Best-UCAM system (p=0.03).

In Table 2.3.7, an overview of the semantic performance across all evaluations is given. The scores are obtained by comparing for each user turn the N-Best list of dialogue act hypotheses, generated by the SLU component, with the reference dialogue act obtained by running the semantic parser on the audio-transcriptions. The used metrics are based on comparisons at either the full dialogue act level, or on the level of the semantic items dialogue acts consist of, i.e., the dialogue act type and the slot-value pairs.

For more details about the metrics used, see [2]. The ASR performance results presented earlier are also reflected in the overall semantic performance (see in particular the scores for SemAcc and ICE[2]), i.e. the scores for the Feb'11 evaluations are much worse than those of the Nov'11 evaluation, the Feb'11 evaluation on Cambridge users getting the worst performance. Furthermore, the oracle vs. top-hypothesis accuracy scores (see for example ODAcc vs. TDAcc) indicate that the N-Best systems had additional information from alternative hypotheses at their disposal to decide on their response actions, but this was only reflected to some extent in task success scores in the Feb'11 evaluation.

Trial	System	OAcc	ODAcc	SemAcc	TAcc	TDAcc	ACE	ICE
Nov'10	N-Best-UCAM	74.2	61.7	65.8	66.9	52.3	3.933	2.077
	1-Best-UCAM	71.0	53.5	70.5	70.5	53.5	4.662	2.216
	N-Best-CLASSiC	75.6	63.9	66.9	68.2	53.7	4.079	1.971
Feb'11	N-Best-UCAM	67.2	59.8	58.1	59.3	49.3	4.428	2.375
	1-Best-UCAM	53.1	37.3	52.2	52.2	37.3	6.591	3.463
	N-Best-CLASSiC	67.6	59.6	58.0	59.1	49.3	4.590	2.394
Feb'11 Mturk	N-Best-UCAM	65.2	55.2	56.1	57.6	46.4	4.559	2.418
	1-Best-UCAM	60.6	44.4	60.1	60.1	44.4	5.563	2.922
	N-Best-CLASSiC	62.7	55.3	53.8	54.8	46.3	4.827	2.569

Table 2.3.7: Semantic evaluation scores, including the item level, confidence weighted semantic accuracy (SemAcc), and the item level and dialogue act level oracle accuracy (OAcc and ODAcc), top-hypothesis accuracy (TAcc and TDAcc), and cross entropy (ICE and ACE) [2].

With the aim of giving some insight as to how the systems perform at different noise levels, plots of predicted success rate against word error rate were generated using logistic regression. Figures 2.3.1 and 2.3.2 show the plots for each system of the subjective resp. objective (assigned goal based) success rate, for the Feb'11 Cambridge user evaluation. Each point in the graph represents one dialogue, unfilled points being successful dialogues and filled points unsuccessful dialogues. For each curve, the standard error is indicated by two dotted lines.

Although the error margins are too substantial for making any strong claims, the plots suggest that the predicted performance of the N-Best-UCAM (CambridgeNBest) system improves relative to that of the 1-Best-UCAM system as the error rate increases.

Figures 2.3.3 and 2.3.4 show the logistic regression plots for the Feb'11 MTurk evaluation. What can be noticed here is that the subjective success rate for the N-Best-CLASSiC system decreases severely as the error rate increases. This might be because at higher error rates, the NLG generating too much information based on incorrect assumptions has a stronger negative impact on perceived success.

2.4 Discussion and Conclusions

The results obtained in the TownInfo evaluations carry some mixed messages. We believe that the decision to switch to a proper telephone-based system on a real-world application domain was the right one, but clearly the lack of training data for the recogniser significantly impacted the recognition performance as shown in Table 2.3.2. In effect, this poor speech recogniser is equivalent to a reasonable speech recogniser operating in very high noise environments. However, even with average WERs over 50%, as shown in Table 2.3.4, subjective success rates of 60% to 65% were achieved in the Feb'11 trial. This shows that

[2]Smaller values of the ACE and ICE metric indicate increasing information content.

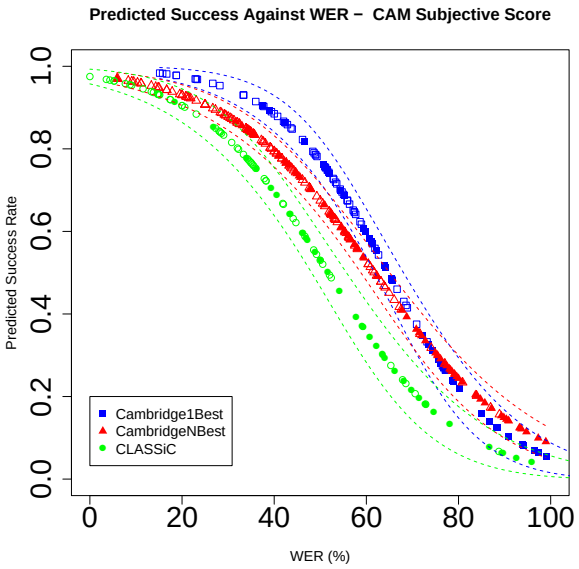

Figure 2.3.1: Logistic regression of *subjective* success rate against word error rate (Feb'11).

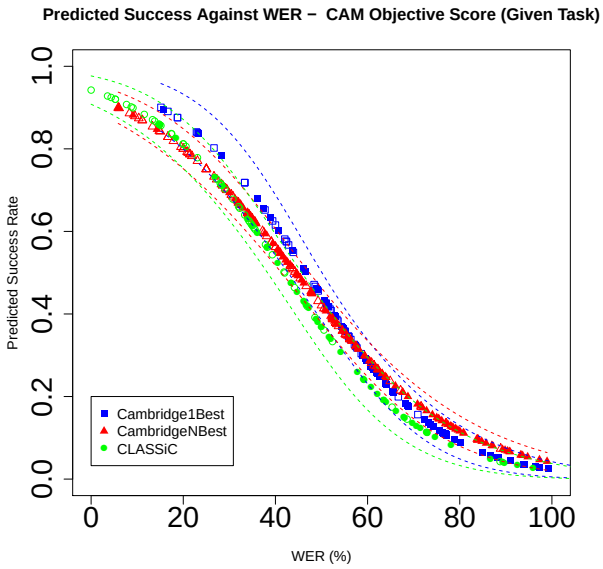

Figure 2.3.2: Logistic regression of *objective* success rate against word error rate (Feb'11).

the systems were fairly resilient even when operating in extremely hostile conditions. It should also be noted that, barring a few initial glitches, all systems operated reliably and robustly during the trials. This was a consequence of the substantial engineering effort put into the overall system development. Also, the improved HIS state space representation and pruning algorithms worked well enabling the systems to support prolonged dialogues without noticeable degradation in real time performance.

Whilst the evaluation results for both the Nov'10 and Feb'11 trials demonstrate the robustness of the systems in severe conditions, the overall performance was rather poor, chiefly we believe due to the poor ASR performance. As noted in Section 2.3.1, when the data collected in the CamInfo trial was used to retrain the recogniser, the word error rate approximately halved and the dialogue success rate increased by over 20%. The on-going trial is testing the N-Best-UCAM system with retrained new acoustic and language models on MTurk recruited subjects. When comparing the perceived success rate, the perceived success rate increased from about 64% to 87% ($p < 0.05$). With this initial experimental result, it is reasonable to expect that all the success rates reported above would similarly increase.

It is also interesting to compare the performance of individual systems. In the Nov '10 and Feb'11 MTurk evaluations, the partial completion score for the CLASSiC system (i.e. including the trained NLG component) was significantly higher than the TownInfo system (p=0.02, z-test), suggesting that the more elaborate venue offers from the trained NLG component helped the user find the venue they were looking for more easily.

However, from figures 2.3.1 and 2.3.3, the CLASSiC system appeared to be more fragile than the other systems at high word error rates, especially in the MTurk subjective evaluation. This might be a consequence of trying to provide too much information to the user based on incorrect assumptions. This suggests that if the system is unsure, it should focus on offering a single entity but when confidence is high, the more intelligent presentation of information generated by the CLASSIC NLG system worked well.

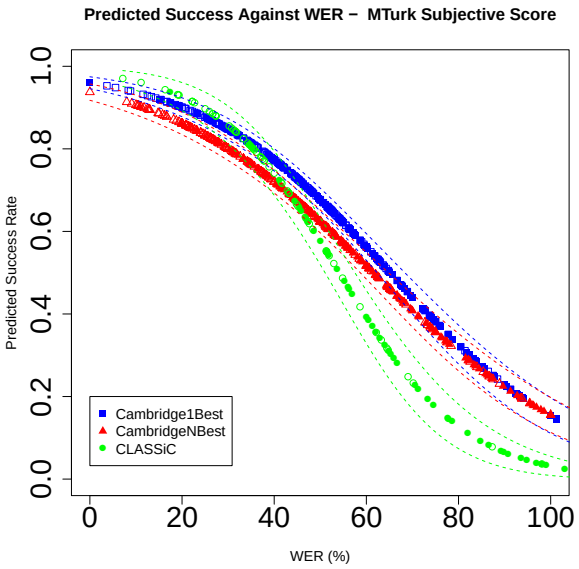

Figure 2.3.3: Logistic regression of *subjective* success rate against word error rate (Feb'11-MTurk).

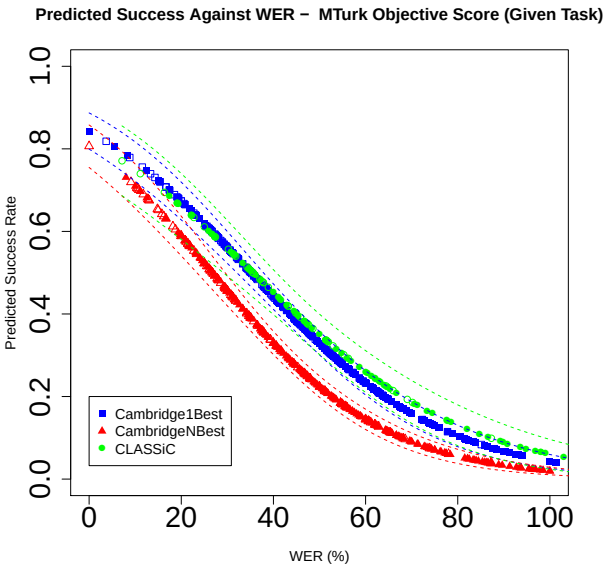

Figure 2.3.4: Logistic regression of *objective* success rate against word error rate (Feb'11-MTurk).

We also note that in one case, the Cambridge based Feb'11 evaluation, the inferred goal based success rates for the N-Best systems are better than those of the 1-Best system, and the success rate for the N-Best-CLASSiC system is significantly higher than that for the 1-Best-UCAM system (p=0.03).

From Table 2.3.4, there is no statistical difference between the 1-Best-UCAM system and the N-Best-UCAM system. This is probably due to a poor match between the simulator error model and real data, which has a greater impact on the N-best policy than on the 1-best policy. This is why it is crucial that we should start training our systems more directly on real user behaviour rather than via handcrafted user and error simulations. Also, in more severe conditions like the Nov'10 and the Feb'11 trials, it is difficult to produce high quality N-best lists as well as confidence scores from ASR and SLU. The lack of additional reliable information in the N-best lists may also have contributed to the similar performance.

However, it is worth emphasising that the 1-best and N-best having similar performance does not mean that the POMDP framework is ineffective, or that an MDP would have worked just as well. In fact, a POMDP system is fundamentally different from an MDP system because the POMDP integrates over both time and the N-best alternatives, whereas an MDP simply tracks the most likely state. Hence, it is not appropriate to draw the conclusion that an MDP would have worked just as well as the POMDP.

Both the 1-best and N-best systems in the trials are HIS systems. As the CamInfo domain is more complex and dialogues with real users tend to be less focused than lab users, the limitation of the HIS framework not being able to support the user changing their goal becomes more of a concern. Also, our intention to extend the summary space was thwarted by the need to develop and understand more sophisticated training approaches (eg GP-SARSA) which can scale to large dimensional state spaces.

A final observation regarding Table 2.3.4 is the significant difference between objective and subjective success rates. This might be partly due to lack of motivation by users in making sure that what they were offered met their requirements and the criterion for objective success might be too stringent. Even though

the use of VoIP technology enabled us to efficiently recruit a large number of users, the nature of paid-for-testing still leads to the pervasive problem of users lacking proper motivation. This has demonstrated yet again the difficulty of testing dialogue systems without having the benefit of a real live application where users call the system because they want to.

Part II

Appointment Scheduling evaluations

Chapter 3

Experiment set-up, system descriptions, and metrics

These experiments explore possible improvements of a service that has been tested and validated during the deployment of a large real-world commercial system: the 1013+ service, which is the Spoken Dialogue System for land line troubleshooting for France. It receives millions of calls a year and schedules around 8000 appointments a month. When the user calls the system, she is presented with an open question asking her for the reason of her call. If her land line is out of service, the Spoken Dialogue System then performs some automated tests on the line, and if the problem is confirmed, it tries to schedule an appointment with the user for a manual intervention. If the system and the user cannot agree on an appointment slot, the call is transferred to a human operator.

This commercial Appointment Scheduling (AS) system was the first large-scale commercial deployment for a spoken dialogue system with on-line reinforcement learning [3]. Systems 2, 3, and 4 are different approaches to the AS problem.

System 2 is the version built using components developed by the academic partners in the project, also using the FT speech synthesiser. System 3 is the adaptation of the commercial system (1013+) developed by FT [4]. System 4 is a lab version of this one, with a list of modifications decreasing the system's influence on the user when interacting [5]. These systems are briefly described in section 3.2. For more information please check the previous CLASSIC deliverables.

This experiments produce evaluations at different levels:

- Statistical comparison of the objective and subjective key performance indicators for each system. This answers the following questions: which system has the best task completion rate? Which system is the more efficient? What are the correlations between the key performance indicators? How do the testers rate the global evaluation of the dialogue for each system? How significant are those results?

- Effects of the question opening on the key performance indicators. It answers the following question: how does this affect task completion? Is the user's satisfaction higher? Is it more efficient? This study is delivered along with an analysis of the regressions that show how the user evaluates the overall rating based on the information that is available on-line and based on the other user questionnaire answers.

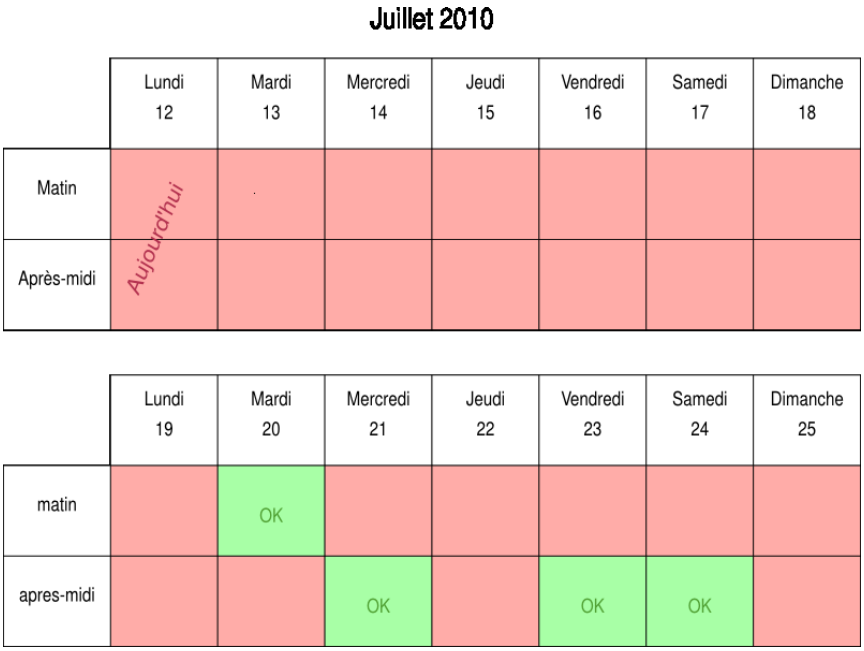

Figure 3.1.1: Example of a user's personal calendar

- A sociological study: identification and description of patterns of user behaviour with the spoken dialogue system, identification and description of types of errors and interactional misalignment phenomena between the user and the system, description of types of interruptions and their contribution to interactional misalignment. On the basis of these three analyses, the sociological report suggests strategies for error recovery. (See Chapter 5)

3.1 Experimental setup

In this domain, the user tries to make an appointment for an engineer to visit their home. Each user is given a set of 2-week calendars which showed their availability and the goal is to arrange an appointment when both they and the engineer are available. An example user calendar is shown in figure 3.1.1.

For the sake of simplicity, we fixed the "today" date to be Monday the 12th of July 2010. It was made clear on the agendas. It was also quite clear that it was impossible to an appointment for "today" and that the appointment was to be made within two weeks. There was no availability on Sundays and the granularity was half a day (a.m. or p.m). Eventually, there were 22 possible appointments and only 1 to satisfy both user and engineer agendas.

There were 12 possible scenarios that were evenly rotated across participants and systems. Each scenario is categorised in terms of scheduling difficulty (Hard/Medium/Easy). Scheduling difficulty is calculated for the User Difficulty (UD) and System Difficulty (SD) separately to assess the system's mixed initiative ability. Scheduling difficulty is calculated as the ordinal of the first session that is free for both the User and the System. Hard scenarios are with an ordinal of 3 or 4; Medium with an ordinal of 2, and Easy with an ordinal of 1. There are 4 scenarios in each of these difficulty categories for both the user and system. To give an example, in Scenario 10, the user can schedule an appointment on Wednesday afternoon but he/she

also has one free session on the previous Tuesday afternoon when the engineer is busy therefore UD = 2. For the system, in this scenario, the first free session it has is on the Wednesday afternoon therefore SD=1. In this case, the scenario is easier for the system than the user because the system could just offer the first session that it has free.

3.1.1 Recruitment method

We recruited in two steps: first we sent a link towards a website for people to subscribe to the experimentation. Several weeks later, we sent each subscriber a list of six hyperlinks for each call. There was a code associated with this hyperlink so that we could make sure each call is unique. This code contained the following information: call identifier (5 digits), called system and the scenario number (2 digits). A last digit was used in order to have a Cyclic Redundancy Check (CRC).

The hyperlink directs the user on another website where the phone number, the code and the user agenda is displayed. The user is asked to call the phone number, to enter the code in DTMF and to try and make an appointment with the system according to the availability in her agenda. When the user calls, a frontend system welcomes her and asks her to enter the code. If the code is erroneous (CRC) or had already been used, the call is denied. If the code is correct, the call is automatically routed to the corresponding system with the engineer agenda that was planned for this scenario.

Each subscriber was supposed to perform six calls but obviously some of them got tired before the end. As the scenarios were totally randomised and as we were not specifically interested in the user follow-up, we decided to keep all the calls in the database. In the end, we collected 628 dialogues for System 2, 740 for System 3 and 709 for System 4.

3.1.2 Questionnaire Content

After each scenario, participants were asked to fill out a questionnaire composed of 11 questions given below. The first three questions were designed to extract information on Task Success (TS) which is either 1 or 0 and measures whether a date was booked that is free for both the engineer and the caller. Objective Task Success is calculated taking Question 3 and comparing it to the calendar for a given scenario. Subjective Task Success is derived from the answer to Question 2. Questions 4-8 are based on the PARADISE USER SATISFACTION questionnaire [6], with one variation where Expected Behaviour is substituted for a question more tuned to extract the quality of the Natural Language Generation component. These questions were on a 6 point Likert Scale so that the users were forced to give either positive/negative feedback and not neutral. Finally, Question 10 is the single RATING metric on a scale of 1-10.

1. Avez-vous obtenu un rendez-vous ? (Have you got an appointment ?)
   Oui /Non/ Je ne sais pas (Yes/ No/ I don't know)

2. Ce rendez-vous correspondait-il à un emplacement libre dans votre emploi du temps ? (Did this appointment match a free timeslot in your schedule ?)
   Oui/ Non /Je ne sais pas [Subjective TS]

3. Quand avez-vous pris votre rendez-vous ? (When is your appointment ?)
   Jour Date Période (Day /Date/ Period) [Objective TS]

4. Lors de votre appel, le système comprenait ce que vous disiez. (During your call, the system understood what you said)
   Tout à fait d'accord / D'accord / Plutôt d'accord / Plutôt pas d'accord / Pas d'accord / Pas du tout d'accord (Completely agree / Agree / Mostly agree / Mostly disagree / Disagree / Completely disagree)[ASR performance]

5. Lors de votre appel, vous compreniez ce que le système disait. (During your call, you understood what the system said)
   Tout à fait d'accord / D'accord / Plutôt d'accord / Plutôt pas d'accord / Pas d'accord / Pas du tout d'accord (Completely agree / Agree / Mostly agree / Mostly disagree / Disagree / Completely disagree) [NLG performance]

6. Lors de ce dialogue, la voix du système était agréable. (During this dialogue, the system voice was pleasant)
   Tout à fait d'accord / D'accord / Plutôt d'accord / Plutôt pas d'accord / Pas d'accord / Pas du tout d'accord (Completely agree / Agree / Mostly agree / Mostly disagree / Disagree / Completely disagree) [TTS performance]

7. Lors de cet appel, vous a-t-il été facile de prendre un rendez-vous ? (During this call, was it easy to make an appointment?)
   Très facile / Facile / Plutôt facile / Plutôt difficile / Difficile / Très difficile (Very easy / Easy / Quite easy / Quite difficult / Difficult / Very difficult) [Task Ease]

8. Dans le cas où vous auriez rencontré des difficultés, veuillez nous expliquer lesquelles. (In case of problems encountered, explain them)

9. Cet appel m'a incité à utiliser un système tel que celui-ci dans le futur. (This call has prompted me to use such systems in the future)
   Tout à fait d'accord / D'accord / Plutôt d'accord / Plutôt pas d'accord / Pas d'accord / Pas du tout d'accord (Completely agree / Agree / Mostly agree / Mostly disagree / Disagree / Completely disagree) [Future Use]

10. Quelle note attribueriez-vous à ce dialogue (note de 1 à 10, 10 étant la note la plus élevée) ? (What grade would you attribute to this dialogue on a 1 to 10 scale, 10 being the best mark)

11. Avez-vous d'autres remarques ou commentaires ? (Do you have other remarks or comments ?)

3.2 System descriptions

As explained earlier, three systems were tested in the CLASSiC AS experiments:

- System 2: HTK ASR and POMDP-based DM developed by University of Cambridge, NLG developed by University of Edinburgh and Heriot-Watt University, and TTS delivered by FT.
- System 3: designed by FT with the off-the-shelf Telisma ASR, it is supposed to be the commercial system direct adaptation to the experimentation set-up. In fact, some modifications were quite significant. Subsection 3.2.2 details them.

- System 4: also designed by FT, it is a lab version of System 3, leaving a wider user initiative in the appointment proposal. This implied to redevelop the whole application, the DM, the prompts (NLG), the ASR with an extended grammar. We also had to We postulate that the System 4 enables the user to interact in a less constrained way (see section 4.4.1).

We now cover each system in more detail.

3.2.1 System 2 description

System 2 deploys integrated components developed by the academic partners, as described in deliverable D5.4. The ASR module uses the Cambridge HTK/ATK speech recogniser, and produces N-best hypotheses with sentence-level confidence scores from speech input. The development of the French ASR system consisted of phone set selection and dictionary construction, acoustic model training and language modelling, using a variety of data resources.

The SLU component makes use of the Phoenix semantic parser and contains a handcrafted grammar to generate dialogue act hypotheses based on the ASR hypotheses. Dialogue acts consist of a type, such as `inform` or `confirm`, and a list of slot-value pairs, such as `dayofweek=Tuesday` or `fromweek=next`. For example, the utterance "jeudi matin" should be decoded as `inform(dayofweek=Thursday,time=am)`.

The dialogue manager takes the N-best list of semantic hypotheses and decides on an appropriate response dialogue act, based on the updated dialogue state. The DM uses the BUDS (Bayesian Update of Dialogue State) POMDP framework, using a Dynamic Bayesian network for monitoring the belief state, and a stochastic policy for deciding on response actions based on that belief state. The policy can be optimised using reinforcement learning, in interaction with an agenda-based simulated user, that was extended for the appointment scheduling domain. In the experiment, both a handcrafted and a learned policy were used.

The natural language generation (NLG) component consists of a rule-based baseline system covering the full range of output system acts, and a specialised component focusing on the generation of temporal referring expressions (see deliverable D5.4). Using reinforcement learning, this specialised component has been optimised for conveying target appointment slots to the user using expressions that are easy to understand, but are not too long, and are preferred by the users.

Two variants of System 2 were evaluated: with and without this trained NLG component (see section 4.2.1).

Finally, the TTS component uses the French Baratinoo expressive speech synthesiser, using a unit-concatenation technique.

For more details on System 2, see deliverable D5.4.

3.2.2 System 3 description

System 3 is the 1013+ commercial system adapted to the experimentation. It involves details such as a more complete logging, a static set-up for the current day: the Monday 12th of July 2010 or the interface with the fake engineer agendas that were designed in the scenarios. But it also required deeper changes which we discuss in this subsection.

The dialogue logic was optimised on-line to the following strategy: first propose the next available time slot to the user. If she/he rejects it, ask her/him for their next availability in the `DAY_OF_THE_WEEK/AM_PM`

format. After a new appointment failure, the commercial system would give up and transfer the caller to a human operator. Instead, System 3 keeps proposing its next availabilities until the end of the calendar. As a conclusion, the dialogue is first system initiative, then user-initiative and finally until the end of the dialogue, system initiative.

Three expressive TTS variants have been used for each dialogue prompt: 'normal', 'calm' and 'dynamic' speaking styles. The goal was to use the experimentation to learn something about the style to adopt in the different contexts. Eventually, no optimisation was found in the experiment and therefore differences between the variants were small. We believe that these variants did not interfere with the other studies and that System 3 can still be used as a baseline for Systems 2 and 4 analyses.

3.2.3 System 4 description

This subsection describes System 4 by enumerating its differences from System 3. The goal was to have a system allowing more user initiative in the appointment proposal. To this objective, System 4 uses dialogue strategies that are much more open than the ones of System 3. As a consequence, the answers from the users are less formatted and it is necessary to extend the ASR grammar. Any grammar extension increases the error rate. This is the reason why, in order to overcome this difficulty, several recovery strategies have been defined. This subsection provides a description of the technical modifications that have been made for System 3, while Section 4.4 explains more in details the consequence implied by these modifications.

Dialogue strategies

As for System 3, the service is organised in rounds of negotiations. In each round, the system attempts to find a new timeslot suitable for the user. Two main strategies are used by the system : either the system directly proposes a timeslot to the user (system initiative strategy), or it asks the user for several pieces of information about either a day, a week or a period during the week she would be available, until being able to propose a timeslot matching the user's constraints (user's strategy).

The first picked strategy is to ask an open question about the user's availabilities: "Quand êtes-vous disponible ?"[1]. Then, the dialogue logic selects the most discriminant criterion given the system agenda and the current user's constraints. This criterion is obtained thanks to a very simple entropy calculation. If there are only 2 or less time slots left, the system switches to the System initiative strategy. It may also switch on this strategy as a result of a user request: "dès que possible"[2] or as an ASR reject recovery strategy.

ASR grammar extension

The ASR module uses the Telisma commercial Telispeech speech recogniser. It gets live speech input from a dedicated telephony board and outputs N-best hypotheses with sentence-level confidence scores. In this final version of System 4, the dictionary is about 200 words: numbers ("le premier", "le dix", ...)[3], months

[1] "When are you available ?"

[2] "as soon as possible"

[3] "On the first", "On the tenth", ...

("juillet", "août", ...)[4], day of the week ("lundi", "mardi", ...)[5], relative expressions for days/weeks ("ce jeudi", "demain", "cette semaine", ...)[6], yes and no ("absolument", "pas du tout", ...)[7], request for repeating ("pardon, je n'ai pas compris", "répète", ...)[8], request for an initiative switch (SYSTEM: Quand êtes vous disponible ? USER: Dès que possible.)[9], ... The user can now express incomplete constraints: "mardi"[10] and formulate complex constraints "la semaine prochaine dans la matinée"[11].

As Telispeech's ASR confidence measure does not deliver probabilities, we had to calibrate and combine the ASR results. The calibration process consists of observing the confidence scores on an annotated corpus and to calibrate the probability of a score being true to the probability that was observed on this corpus. The combination process consists of considering each ASR result in the n-best list as an independent probabilised piece of information and combining their probabilities in order to compute a probability distribution on the n-best list. In addition to providing reliable probabilities, it reduces the false acceptance / false rejection trade-off by 40%.

ASR reject recovery strategies

Sometimes, ASR returns a reject for an audio input. It either means that the confidence level for the best transcription is too low or that nothing in the language model has fit the input. Such rejects are common and we have experienced in CLASSIC System 4 five alternatives to deal with such errors :

- Feedback + repetition : the system informs the user that it is not sure of what it has heard and then it repeats exactly the same question. This is somehow the baseline for the study, as implemented in CLASSIC System 3.
- Feedback + energetic prompt : as well, the system feedbacks the reject but instead of repeating the same .wav file, we use a more energetic TTS variant that stresses important parts of the question.
- Feedback + rephrasing : the system still feedbacks but rephrases instead the question in another way.
- Feedback + yes/no question : if the ASR's 1-best transcription is not too low, the system asks the question : "Avez-vous dit *1-best* ? Merci de confirmer par oui ou par non."
- Feedback + change of strategy : the system still feedbacks the reject, but it switches to a system initiative strategy (as described in the previous subsection).

Each user's n-best answer is submitted to a confidence test based on the ASR acoustics score, and its uncertainty is integrated into the Context Manager [7]. For the confirmation question, the distribution of answers output by the ASR module are used to shift the ASR confidence level score of the 1-best and give better precision and recall performances.

[4] "July", "August", ...
[5] "on Monday", "Tuesday", ...
[6] "this Thursday, tomorrow, this week, ..."
[7] "sure", "not at all", ...
[8] "sorry, I did not understand", "repeat", ...
[9] SYSTEM: When are you available ? USER: As soon as possible
[10] "on Tuesday"
[11] "Next week in the morning"

3.3 Evaluation metrics

This section describes the key performance indicators that were collected during the experimentation. Firstly, subsection 3.3.1 details those key performance indicators. Secondly, subsection 3.3.2 explains how the statistics reports of appendix A should be read.

3.3.1 Collected key performance indicators

The key performance indicators are restricted to the ones that could be gathered automatically from the system logs or from the user questionnaires. The key performance indicators that were considered are the following ones:

- *Sys. task completion* is the objective task completion: did the system book the good appointment regarding the scenario?
- *Q. task completion* is the subjective task completion: did the tester think she had booked an available slot?
- *Call duration*: how long (in seconds) did the dialogue last?
- *Number of ASR rejects*: how many times per dialogue did the system acknowledge not understanding the tester's utterance?
- *Q. ASR rating*: how good did the tester rate the understanding of the system? (6 is the maximum)
- *Q. phrasing rating*: how good did the tester rate the understandability of the system? (6 is the maximum)
- *Q. TTS rating*: how good did the tester rate the pleasantness of the system? (6 is the maximum)
- *Q. overall rating*: how good in general did the tester rate the quality of the dialogue? (10 is the maximum)
- *Q. task ease*: how easy did the tester rate the task? (6 is the maximum)
- *Q. future use*: would the tester use such a system? (6 is the maximum)
- *Number of call*: was this call the first, second, third, fourth, fifth or sixth of the user in the experimentation?
- *Volume*: how many calls were collected?

3.3.2 Explanation of the statistics presented

As explained above, there were 12 scenarios that were dispatched into a user-difficulty and a system-difficulty classification. In order to show the correlation of the task difficulties over the key performance indicator, we displayed a 4x4 grid for each statistic of interest (see Figure 3.3.1). The 4x4 grid is read as follows:

Overall	S-diff = 1	S-diff = 2	S-diff = 3
U-diff = 1	U1 & S1	U1 & S2	U1 & S3
U-diff = 2	U2 & S1	U2 & S2	U2 & S3
U-diff = 3	U3 & S1	U3 & S2	U3 & S3

Figure 3.3.1: Explanation of the statistics tables.

- In the top-left corner, there are the complete statistics regardless of the user- or system- difficulties
- On the top line (except top-left corner), there are the statistics for system-difficulties (1 is easy, 2 is medium and 3 is hard), regardless of the user-difficulties.
- In the left column (except top-left corner), there are the statistics for user-difficulties (1 is easy, 2 is medium and 3 is hard), regardless of the system-difficulties.
- In the bottom-right colourless 3x3 table, there are the statistics for specific user-difficulties and system-difficulties. For instance, the U2 & S3 case means that the user-difficulty is medium and the system-difficulty is hard.

As an example and a preamble analysis, let us have a look at the volume of collected data (last column) in Figure 1 (see Appendix A). Note that for each system (and therefore also on aggregate), we collected with the user-easy/system-easy, user-medium/system-difficulty and user-difficult/system-medium difficulties twice as many dialogue than with the other difficulties. The reason is that we used 12 scenarios, 2 for the user-easy/system-easy, user-medium/system-difficulty and user-difficult/system-medium difficulties, but only 1 for the other difficulties. The results are fully presented in Appendix A.

Chapter 4

Statistical Analyses

4.1 Evaluation of Systems 2, 3, and 4

4.1.1 Objective evaluation

This section describes, compares and explains for each system the statistics obtained on the objective key performance indicators (Sys. task completion, Call duration, Number of ASR rejects). Please refer to Appendix A for detailed tables of the results.

Objective task completion

For task-based applications, the most important one is the system task completion. Systems 2, 3 and 4 achieved respectively $79.1\% \pm 3.2\%$, $80.8\% \pm 2.9\%$ and $82.9\% \pm 2.8\%$. The first remark is that all those systems performed very well compared to the commercial system task completion rate which is between 70% and 75%. The only significant difference between the system task completion rates are between Systems 2 and 4. Please recall that System 2 had a different speech recogniser to System 3 and System 4, and so any comparisons must take this fact into account.
Remarks on the effect of the scenario difficulties on the system task completion rates for each system:

- System 2 performs worse than the others for the scenario that is both user and system-difficult.
- System 3 performs worse when the scenario is system-difficult and not user-easy, which is logical if you remember its dialogue strategy: first system initiative, then user initiative, then system initiative until the end of the dialogue.
- System 4's task completion rate does depend on the scenario's system-difficulty but strongly and significantly on the scenario's user-difficulty, which was also predictable since System 4 is almost exclusively user-initiative.

Call duration

Dialogues with System 3 are significantly shorter. We explain this by the fact that this system controls better the user's answers, with a lot of help for the tester to format his/her answers. Moreover, as the

system initiative is the main strategy, most of the questions expect a yes/no answer. As a consequence, there are less ASR rejects and errors (see next paragraph) and therefore dialogue turns.
Remarks on the effect of the scenario difficulties on the call duration for each system:

- System 2 call duration looks surprisingly more dependent on system-difficulty than user-difficulty.
- As predicted, System 3 call duration is not dependent on user-difficulty but highly dependent on system-difficulty.
- For System 4, it is the opposite, which was also predictable.

Number of ASR rejects

Systems 3 and 4 are built in such a way that System 4 ASR rejects almost five times more than System 3 ASR. Indeed, System 3 constrains the user answers while System 4 expects a wide range of answers and had also its confidence score threshold raised to ensure to avoid as much as possible ASR false acceptances[1].
Remarks on the effect of the scenario difficulties on the number of ASR rejects for each system:

- We expected to have a strong correlation between the average call duration and the number of ASR rejects. However, the System 3 number of ASR rejects is strongly dependent on user-difficulty.

Please note that a more detailed analysis of the System 2 results is given in section 4.2.

4.1.2 Subjective Evaluation

This subsection describes, compares and explains for each system the statistics obtained on the subjective key performance indicators (Q. task completion, Q. ASR rating, Q.phrasing rating, Q. TTS rating, Q. overall rating, Q. task ease, Q. future use). Please refer to figure 1 for details.

Subjective task completion

The user was asked two successive questions: “did you book an appointment?” and “Was this appointment available?”. The subjective task completion rate is set to 1 if answered yes to both questions. This is the key performance indicator that informs the task completion as perceived by the user. System 2 does not score as well on this key performance indicator ($68.2\% \pm 3.7\%$). It seems that some testers were puzzled and could not tell whether their appointment was booked. The low correlation on Figure 3 between the system and subjective task completion rates confirms it. At the opposite, Systems 3 and 4 testers tended to slightly overrate their performance with respectfully $83.0\% \pm 2.8\%$ and $85.0\% \pm 2.7\%$ subjective task completion rates.
Remarks on the effect of the scenario difficulties on the questionnaire task completion rates for each system:

- For Systems 3 and 4, the correlation of the scenario difficulty with the subjective task completion rate is less obvious than with system task completion rate.

[1] The disambiguation strategy was counted as a reject, although it is only half-rejected.

Subjective ASR rating

System 3, once more, takes benefit from its simplicity. The scores for Systems 2 and 4 that are less constraining of the user expression are lower. System 4's subjective ASR rating is significantly better than that of System 2. However, the difference remains small.
Remarks on the effect of the scenario difficulties on the subjective ASR rating for each system:

- As a general remark, there is a significant correlation between the scenario difficulty and the subjective ASR rating.

Subjective phrasing rating

Systems 3 and 4 have very high subjective phrasing rating. The significant difference between them is probably explained by the fact that the dialogues with System 3 were shorter and therefore less boring. System 2's subjective phrasing rating is far lower, but still at a good level. There is probably a correspondence between this lower phrasing rating and the lower subjective task completion rate.
Remarks on the effect of the scenario difficulties on the subjective phrasing rating for each system:

- Surprisingly, there is no correlation this time between task difficulty and this rating for System 2.
- However, we still find it (but less visible) for Systems 3 and 4.

Subjective TTS rating

All the systems use the same TTS technology: Baratinoo delivered by FT. However, there are some differences :

- System 2 uses the male voice Loïc, while Systems 3 and 4 use the female voice Julie.
- System 2 synthesizes the prompts on the fly, while Systems 3 and 4 used fine-tuned TTS recordings for static prompts (and on the fly TTS for dynamic prompts)
- System 3 implemented variations with three speaking styles for each prompt: neutral, dynamic and calm. The goal to eventually converge to the best style for each context, but no trend was found so that the learning algorithm kept exploring and randomly picked the speaking style.
- System 4 implemented an error recovery strategy by repeating the question with the dynamic speaking style.

System 2's TTS rating is significantly lower than for other systems. The difference between systems 3 and 4 is small but significant. As for the subjective phrasing rating, we believe that this difference is a side effect of the boredom induced by the longer calls with System 4.
Remarks on the effect of the scenario difficulties on the subjective TTS rating for each system:

- As for phrasing rating, there is no correlation this time between task difficulty and this rating for System 2.
- However, we still find it (but even less visible) for Systems 3 and 4.

Overall rating

System 3 is rated the best, which is surprising since the best task completion rates are obtained by System 4. It shows that, for testers[2], the task completion rate is not the main criterion for rating how well a dialogue went. Section 4.4 analyses extensively the correlations between the various ratings of the users. System 2 receives once more a lower rating, paying for both a low subjective task completion and a low subjective ASR rating.

Remarks on the effect of the scenario difficulties on the overall rating for each system:

- Contrarily to the subjective task completion rate, the overall rating is significantly correlated to the scenario difficulties. It is surprisingly close from the system task completion correlation to the scenario difficulties.

Subjective task ease

The task ease question was presented to the tester only when the tester considered he/she completed the task. It was set to 0 in the other cases. It is interesting that the task ease rating is very close to the subjective ASR rating.

Future use

The future use question was presented to the tester only when the tester considered he/she completed the task. It was set to 0 in the other cases. It is interesting that the future use rating is very close to the subjective ASR rating.

4.2 System 2 detailed evaluation

Here we present a more detailed evaluation of CLASSiC System 2, which had two variants: with and without the trained NLG component for Temporal Referring Expressions (see D5.4). 628 dialogues were collected in evaluating System 2.

System 2 achieved an objective Task Success approaching 80%, despite being developed rapidly, and deployed following minimal testing. Here we focus on the evaluation of the NLG component in System (section 4.2.1), and on a PARADISE-style analysis of the results in section 4.2.2.

We note that System 2 cannot be directly compared with System 3 and System 4, since different speech recognisers were used. However, we can draw some general lessons and conclusions about the performance of the CLASSiC systems and the methods used to develop them (see section 4.3).

4.2.1 System 2 NLG Results

During this evaluation, we compared two types of NLG methods for generating Temporal Referring Expressions (TRE) for appointment dates. A data-driven policy developed Heriot-Watt and Edinburgh (see Deliverable D5.4 and [20]) was activated when the system informs the user of an available time slot. This

[2]It is probably different for real users.

system was compared to the exact same system but with a *rule-based* adaptive baseline system (developed at the University of Cambridge). In this rule-based policy MONTH, DATE and TIME were always absolute, DAY was relative if the target date was less than three days away (i.e. "today, tomorrow, day after tomorrow"), and WEEK was always relative (i.e. "this week, next week"). All 5 information units were included in the realisation (e.g. "Thursday the 15th July in the afternoon, next week") although the order was slightly different (DAY-DATE-MONTH-TIME-WEEK).

Parameters	**Learned TRE**	**Baseline TRE**
Actual Task Success	80.05%	78.57%
Perceived Task Success	**74.86**%*	60.50%
User satisfaction	**4.51***	4.30
No. system turns	22.8	23.2
Words per system turn	**13.16***	17.3
Call duration (seconds)	**88.60** *	105.11

Table 4.2.1: System 2 NLG evaluation: Results with real users (* = statistically significant differences at $p<0.05$)

Results from the System 2 NLG study are summarised in Table 4.2.1. The data-driven NLG policy ('Learned TRE') showed significant improvement in Perceived Task Success (+23.7%) although no significant difference was observed between the two systems in terms of Actual Task Success (Chi-square test, df=1). Overall user satisfaction (the average score of all the questions) was also significantly higher (+5%)[2]. Dialogues with the learned policy were significantly shorter with lower Call Duration in terms of time (-15.7%)[2] and average words per system turn (-23.93%)[2].

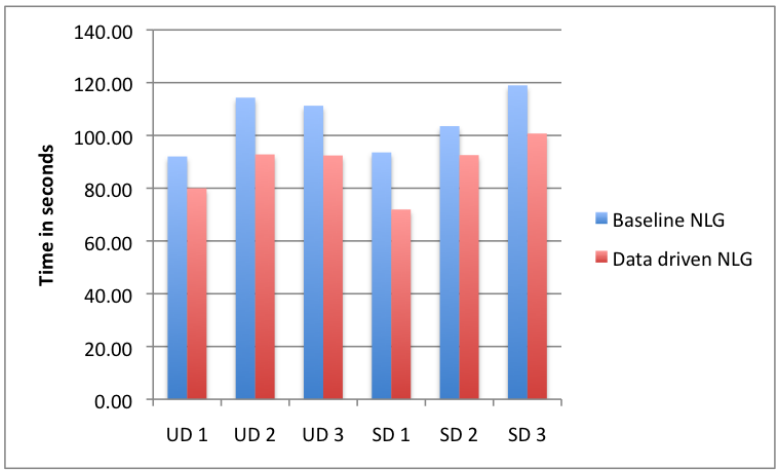

Figure 4.2.1: Graph comparing length of dialogues for user (UD) and system difficulty (SD)

Figure 4.2.1 shows the length results in time for systems of varying UD and SD. We can see that the data-driven adaptive policy consistently results in a shorter dialogue across all levels of difficulty. In

summary, these results show that using an NLG policy trained on data results in shorter dialogues and greater confidence in the user that they have had a successful dialogue.
These results are to be published in [20].

4.2.2 System 2 Results: PARADISE-style evaluation

We applied the PARADISE framework [6] to develop models of both USER SATISFACTION and RATING as the response variable. The entire dataset was used to build each model. PARADISE is an evaluation framework that uses USER SATISFACTION, in our case, calculated by summing the answers to Questions 4-8 in the questionnaire above. This USER SATISFACTION metric is taken as the overall objective to be maximized by a system, and then task success and various interaction costs can be used as predictors of USER SATISFACTION. PARADISE uses multiple linear regression to quantify the relative contribution of these predictors reflected in their coefficients.
[8] and others argue that it is over-ambitious to directly relate a single metric such as RATING to a measure of overall system quality. Rather it is better to limit the scope of the perception and judgment component to the prediction of values on a number of perceptual quality dimensions. Obtaining values on these dimensions, however, requires administering lengthy questionnaires, which is not necessarily plausible for real-user evaluations "in the wild". Therefore, here we investigate what information we can gather on the various components by performing multiple linear regression using a single RATING from Question 10 as the response variable.

Metrics collected per call are divided into Dialogue Efficiency, Dialogue Quality, Task Success, and User Satisfaction in line with the PARADISE framework:

- **Dialogue Efficiency**:Task duration (in seconds), system turns, user turns, total turns, av words per user turn, av words per system turn,
- **Dialogue Quality**: Word Error rate (WER), number of ASR rejects,
- **Task Success**: subject and objective measures,
- **User Satisfaction**; sum of TTS/ASR/NLG performance, Task ease, and Future Use (out of 30); single RATING (out of 10).

Table 4.2.2 gives details of the model for USER SATISFACTION where $R^2 = 24\%$ with the coefficients indicating both the magnitude and whether the variable is a positive or a negative predictor of USER SATISFACTION. Here Objective Task Success is the main predictor with a coefficient of 1.95. Creating a separate model with only Objective Task Success as a predictor with no other metrics results in coverage of $R^2 = 5\%$ so clearly the other metrics are also contributing. One is not able to read too much into the predictors as the coverage of the response variable is rather low. However, one can infer that the presence of the system act "Bye" has a positive impact because it indicates that the user has at least arranged an appointment (even if it is not necessarily correct) rather than the user hanging up. "Reqmore" (Request More Information) has a negative coefficient as it indicates that a misrecognition has occurred.
The model of RATING is given in Table 4.2.3 with a coverage of the variance of $R^2 = 36\%$ which is comparable to the initial DARPA Communicator evaluation [9]. Here we can see the clear role of both Objective

[2] independent two-tailed t-test $p < 0.05$

Metrics	Coefficient	P-Value
ObjTaskSuccess	1.95	0.01
Bye	1.21	0.33
Reqmore	-1.00	0.21
Confirm	-0.73	0.29
Request	-0.61	0.37
Inform	-0.57	0.40

Table 4.2.2: Predictive power of core metrics and significance for USER SATISFACTION, $R^2 = 24\%$

and Subjective Task Success. This is a better fit than the one trained to optimize USER SATISFACTION with an increase of 12% variance coverage. Again, we trained a separate model using only Objective Task Success as a predictor, resulting in $R^2 = 20\%$ indicating that the other metrics count for 16% of the remaining variance.

Metrics	Coefficient	P-Value
ObjTaskSuccess	1.57	0.00
SubjTaskSuccess	1.24	0.01
Bye	0.44	0.55
Total Sys Turns	-0.3	0.25
Reqmore	-0.27	056
Total User Turns	0.22	0.64

Table 4.2.3: Predictive power of core metrics and significance for RATING, $R^2 = 36\%$

The second most influential factor in the model given in Table 4.2.3 is Subjective Task Success. In order to obtain this metric, the user must be asked a question such as "did you get the appointment you needed?". Again there may not be time and the user might not be motivated to answer such a question and simply hang up. Taking this into account, we created a model with only the Objective Task Success that can be automatically calculated from the log files along with other metrics. This model, given in Table 4.2.4, results in a drop of only 1% variance compared to the model that uses both types of Task Success. Indeed, the correlation between Objective and Subjective Task Success is 0.81 and when Subjective Task Success is removed the model relies more on dialogue duration predictors and various dialogue acts.

Given previous spoken dialogue system evaluations [9], it is somewhat surprising that the Word Error Rate is not used as a predictor of dialogue quality. This may be because in the domain of appointment scheduling, the user is mostly restricted to responding with preferred dates and yes/no answers. Figure 4.2.2 gives the average WER for Objective and Subjective Task Success. One can see that for Objective Task Success when WER is higher, Task Success is lower on average, i.e. fewer users accomplish the task if WER is higher. However, this is not the case for Subjective Task Success bringing into question the validity of Subjective Task Success as a performance metric.

Figure 4.2.3 gives RATING and USER SIMULATION (normalised to be on the same scale as RATING of 1-10) for the different scenarios of varying Scenario User Difficulty. A graph for Scenario System Difficulty is not shown here but is very similar. This chart illustrates a direct negative correlation of

Metrics	Coefficient	P-Value
ObjTaskSuccess	2.57	0.00
Total Sys Turns	-0.33	0.22
Bye	0.34	0.64
Reqmore	-0.27	0.56
Confirm	-0.20	0.62

Table 4.2.4: Predictive power of core metrics and significance for RATING without Subjective TS, $R^2 = 35\%$

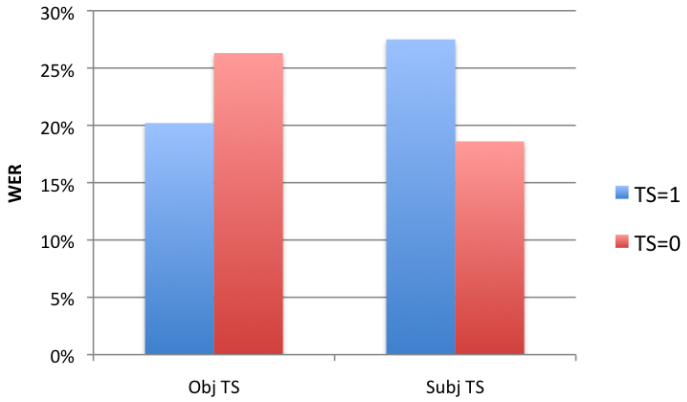

Figure 4.2.2: WER for Objective and Subjective Task Success (TS): TS=1 if successful, = 0 otherwise

RATING and Scenario User Difficulty (-0.997) and slightly less so for USER SATISFACTION and Scenario User Difficulty (-0.91). It is not surprising then that the more difficult the task, the higher the WER and the less likely the user is going to give a high quality rating.

4.3 Summary: Comparing Systems 2, 3, and 4

First, note that comparing Systems 2, 3 and 4 directly is not possible due to the different speech recognition components used. However, we can draw some general conclusions about the comparative performance of the different systems.

Note that commercial systems are typically deployed only after many iterations of user testing. In this case, both System 2 and System 4 were trialled following minimal testing.

Table 4.3.1 shows a comparison between the 3 systems. Here, one can see that actual task completion (Obj TS) is comparable across Systems.

Even though objective evaluations of the three systems are very close, the subjective evaluation is slightly more critical regarding System 2. The subjective task completion rate and the phrasing rating indicate that some of the testers sometimes felt lost about whether their appointment was booked. The TTS was also

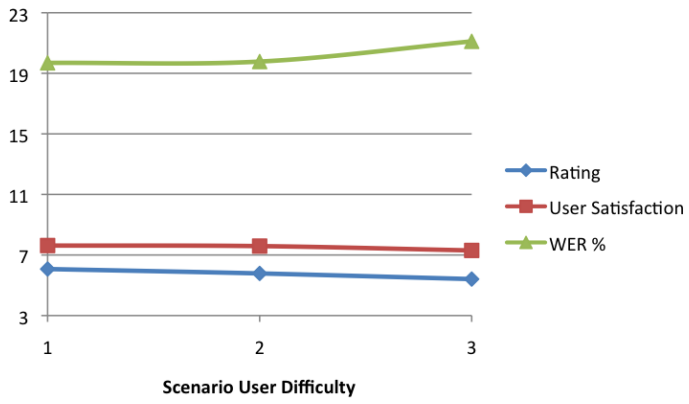

Figure 4.2.3: RATING and USER SATISFACTION (both on a scale of 1-10) and WER for user scenario difficulty

	Obj TS	Subj TS	Av. User Sat	Rating	Call Time (s)	Num Dial
System 2	79.23%	69.09%	4.42	5.29	94.95	605
System 3	81.32%	83.52%*	5.16*	7.44*	68.54*	728
System 4	83.17%*	85.47%*	4.41	6.56*	97.95	695

Table 4.3.1: Systems 2, 3, and 4: Results with real users (*=statistically significant difference with System 2 at $p<0.05$) [Notes: Obj / Subj TS = Objective / Subjective Task Success. Num Dial = number of dialogues. User Sat = User Satisfaction (Max=6), Rating max =10.]

rated lower, which demonstrates how important the TTS handcrafted fine tuning is.

The comparison between Systems 3 and 4 is more complex. System 4 has better task completion rates (objective and subjective) but System 3 is more efficient and pleasant to use: with shorter dialogues and better questionnaire ratings. The divergence between those two systems is explained by a difference of approach: System 3 intends to constrain the user into a very predictable behaviour, while System 4 asks more open-ended questions and endeavours to anticipate every possible user utterance. We feel that the impact of this divergence of approach deserves a more extensive analysis in section 4.4.

Overall, then, System 2 is of comparable quality to the industrial deployed system (System 3) and its more advanced counterpart (System 4), despite the problems arising from the high word-error rates encountered by System 2. We recall that improving ASR performance was not the subject of the project. System 2 was built by a small team using the statistical methods developed in the CLASSiC project, over a period not exceeding 9 months in the final project year. This result shows that the statistical learning methods developed in the project provide a promising foundation for future research and development into robust and adaptive spoken dialogue systems.

4.4 Do users appreciate being constrained?

This section will focus on the comparison of the evaluation of the two FT systems (3 and 4) in order to investigate how the predictability/constraint trade-off is perceived by users and in the end in order to draw recommendations on how to optimize this for the best dialogue performance (namely task completion and user satisfaction). The quality of these systems' handcrafting is grounded by the fact that they got respectively a 80.8% and 82.9% task completion rate, versus 79.1% for System 2.

4.4.1 Qualitative analysis

There are two approaches to handcrafted spoken dialogue system design:

- Usage formatting: the designer constrains the user into predefined behaviours and thus limits the range of his/her expressiveness. This is the approach followed by System 3 with prompts like this one: "Dans ce cas, merci de me dire un jour de la semaine où vous seriez présent à votre domicile, du lundi au samedi, en précisant le matin ou l'après midi. Par exemple : mardi matin. C'est à vous !"[3].

- Usage anticipation: the designer endeavours to anticipate every possible unconstrained user behaviour in order to be able to interpret it and react accordingly. This is the approach followed by System 4 with prompts like this one: "Quand souhaitez-vous prendre rendez-vous ?"[4].

A lot of words could be used in place of "unconstrained": spontaneous, natural or free. FT thinks that none of them is fully satisfying. It is "natural" for a user to answer a yes/no question with "yes" and "no". This is what she would "spontaneously" do. And there is no way to remove the user's "freedom" of speech, just a possibility to influence his behaviour in a cooperative way. The user is still "free" of being non-cooperative. Even the word "constrained" is not perfect, since the user does not necessarily feel "constrained". "Directed" or "influenced" would perhaps be more faithful, but we wanted to keep the slightly pejorative "constrained" word to break the cliché.

For all practical purposes, it is impossible to exhaustively anticipate all the potential user reactions. This approach obviously requires a widening of the ASR grammar: the system needs to be able to understand what the user would spontaneously say after such an open question. However, the impact cannot generally be reduced to this point. New words refer to new concepts (SLU extension) which require new dialogue strategies (DM extension). Concerning NLG and TTS, they usually call for a simple adaptation, but sometimes, the new strategies are so complex that new NLG, and maybe TTS techniques must be used. As a first conclusion, the whole dialogue chain is impacted by such a shift in the approach. Even further, traditional approaches to dialogue assume a "Walkie-Talkie" complex: when the system speaks, it does not listen to the user. And conversely, when the user speaks, the system prevents itself from speaking. However, people do not take turns to talk in a typical interaction. Usage anticipation in the active listening and in synchrony urges to readdress the whole dialogue architecture. This is what incremental dialogue systems [10, 11] intend to do.

[3] "In this case, please tell me a day of the week when you are at your home, from Monday to Saturday, by specifying the morning or the afternoon. For example Tuesday morning. You can speak!"

[4] "When do you want to book your appointment?"

Of course, a spoken dialogue system design is always a balance between those two approaches. Indeed, it is both impossible to constrain perfectly the user into a predefined behaviour[5] or to let the dialogue completely open to everything a user might say[6]. Systems 3 and 4 are two implementations of the same service with respectfully a usage formatting oriented approach and a usage anticipation oriented approach. To make things simple, we say that System 4 is a version of System 3 with a user constraint decrease.

As extensively explained in section 3.2.3, a lot of attention has been turned to keep the task completion rate of System 4 at a high level. Section 4 confirms that our goal has been met. However, section 4.1.2 shows that the users did not enjoy the constraint decrease. Indeed, even recovered, an interactional misalignment leaves the tester with a bad experience and the subjective ratings are impacted consequently.

On the one hand, System 3 constrains the user into a predefined behaviour with long sentences. Here the user knows how to reach his/her goal, but on the other hand, the user may be frustrated to be confined into a strict lengthy process without any possible shortcut.

System 4 lets the user be unconstrained. On the one hand, the user feels comfortable in expressing himself in his specific way but on the other hand, it is made possible at the expense of a lot of ASR confirmation and some ASR rejects or errors. The statistics in section 4 show that System 4 dialogues are 50% longer than System 3 ones on average. This is a direct consequence of this drawback.

The experimental results are clear: testers did not appreciate the question opening proposed by System 4, even if the task completion was better (not significantly though). Even if the ASR errors and rejects did not compromise the task completion, they frustrated the testers and reduced the quality of service. The testers did not perceive the call as an entertainment, but as a task that should be done efficiently. The testers want to be helped to accomplish the task and to be directed as much as possible in the most efficient way. They need to know exactly what the system expects from them even in very intuitive tasks as Appointment Scheduling. And they also expect the system to choose the dialogue strategy that avoids the best the ASR errors and more generally interactional misalignments.

4.4.2 Regressions

This subsection will show and discuss two kinds of regressions that have been performed in order to predict the overall rating:

- Based on key performance indicators that are available on-line: automatically computed task completion[7], dialogue duration and number of ASR rejects[8] per minute of dialogue. The goal of this regression is to understand how well the user satisfaction can be estimated from on-line data.
- Based on user questionnaire key performance indicators: perceived task completion, ASR rating, NLG rating and TTS rating. The goal of this regression is to understand what parts of the system matter for the user satisfaction.

[5] Would it be just because users are not always attentive to what the systems says.

[6] For instance, the system asks "Are you available on Tuesday morning?" and the user might answer "Errr, actually I have to go to see my daughter at 9 a.m., so it depends if it is after 10. But not after 12! Because I don't want to miss my TV show."

[7] This key performance indicator can be considered available for commercial systems too. Indeed, there is always a need for a yes/no appointment confirmation question that is reliable enough to be used as a automatic task completion evaluation.

[8] It includes rejects and time-outs but not ASR errors, which cannot be automatically estimated

On-line regression

Systems 3 and 4 obtained consistent results that still enlighten some differences between the two systems: System 3 user overall rating is almost twice as sensitive to ASR rejects than System 4. This can be explained by the fact that it was quite rare to get an ASR reject with System 3 (see Figure 1). Therefore the obtained numbers for *NB rejects/Duration* were much smaller and it is necessary to have a bigger coefficient to take into account this indicator. The reflexive observation can be made regarding the duration but the explanation is completely different. System 3 had a very organized dialogue strategy, which was simple to apprehend and to use. System 4 was a bit more messy and some users could feel lost (see section 5), which would imply both longer dialogues and a bad user experience. Thus, we interpret this coefficient difference as being caused by this phenomenon. The adjusted R^2 is around 33% for all three

Key performance indicator	Sys. 3	Sys. 4	Sys. 3+4
Intercept	7.23	7.53	7.58
Objective Task Completion	1.26	0.99	1.07
Duration (in minutes)	-0.43	-0.81	-0.74
Nb Rejects/Duration	-0.76	-0.44	-0.60
Adjusted R^2	31.98%	33.00%	34.06%

Table 4.4.1: Questionnaire based regression for predicting overall rating.

dialogue sets. This is very low, since it means that 67% of the variance remains unexplained with our model. Nevertheless, we are used to have such poor user satisfaction models with dialogue systems.

It is interesting to see that a classical 1 minute successful dialogue with 1 ASR reject is rated the same with Systems 3 and 4 (respectfully 7.30 and 7.27). It means that the dialogue approach taken by the systems (usage formatting versus usage anticipation) is not directly the cause of the overall rating difference. It is transparent to the user. However, the dialogue approach implied great differences in the objective key performance indicators expectations, which deeply influenced the overall rating.

Questionnaire-based regression

The questionnaire-based regression was applied to all 3 systems. All the key performance indicators were linearly projected on the -1/+1 segment. For the perceived task completion, a success was set to +1 and a failure to -1. For the other key performance indicators, they were all comprised between 1 and 6, so that the following transformation was applied : $x^* = \frac{x-3.5}{2.5}$.

Table 4.4.2 shows that Systems 3 and 4 regressions are very close – almost similar. We can conclude that the unconscious user rule for selecting the overall rating was not influenced by the user constraint decrease.

The System 2 regression is worse (only 56% for adjusted R^2 versus around 67% for Systems 3 and 4) and quote different. We explain these variations by the fact that testers' populations were not identical. Indeed, most Systems 3 and 4 testers were FT employees while most System 2 testers were Supelec students. Moreover, FT employees were calling from their office (or from their home for a few of them), while Supelec students were generally calling from their classroom.

Whatever the system, it is surprising to see that the ASR rating was at least twice as important as the perceived task completion. We suspect that real users would be more interested in the task completion

Key performance indicator	Sys. 2	Sys. 3	Sys. 4	Sys. 3+4	Sys. 2+3+4
Intercept	3.50	4.27	4.07	4.13	3.75
Perceived Task Completion	1.07	0.87	0.89	0.86	0.97
ASR rating	2.14	2.27	2.36	2.35	2.29
NLG rating	0.96	0.37	0.48	0.46	0.77
TTS rating	0.52	1.21	1.04	1.12	1.04
Adjusted R^2	56.08%	67.24%	66.54%	67.91%	66.04%

Table 4.4.2: Questionnaire-based regression for predicting overall rating.

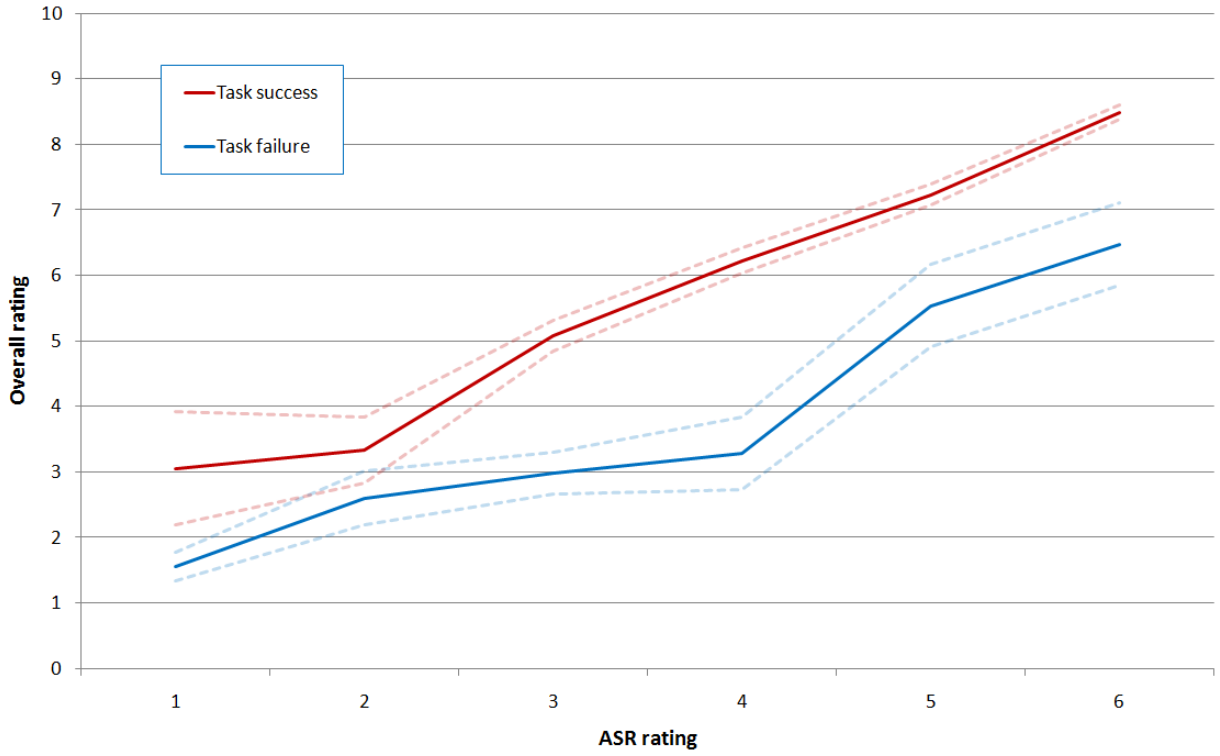

Figure 4.4.1: Overall rating in function of ASR rating when the task completion is perceived as successful or unsuccessful (based on the data of all 3 AS systems).

than the experimental testers were. Concerning Systems 3 and 4, the perceived task completion influence is lower than TTS rating influence. It shows that the pleasantness of the dialogue system was regarded by the testers as more important than its performance.

In order to illustrate the previous finding, Figure 4.4.1 shows graphically how strongly the ASR rating influenced the overall rating. The dotted curves shows the 95% confidence window. The window is narrower when the data is more voluminous.

We now turn to a further detailed qualitative analysis of the AS dialogues using methods from Conversation Analysis, for example examining types of errors and interactional misalignment phenomena between the user and the system. This leads to suggestions of strategies for error recovery.

Chapter 5

Sociological Evaluation Report

This chapter presents the results of the evaluation of users' interactions with the 1013+ automated dialogue appointment scheduling system carried out at FT. 1013+ is a spoken dialogue system based on reinforcement learning that aims to improve the task completion over time by on-line optimisation of the system's design.

Two corpora from two different settings are examined. The first, which we call naturalistic (or commercial), comes from a large-scale commercial deployment System 3 (June 2010). It is composed of 235 user-system dialogues. The second one is experimental and is based on users' trials in which users follow scenarios in an experimental setting (February 2011). System 3 was tested together with System 4, a lab version. Our experimental corpus is made up of nearly 400 dialogues selected from the originally collected dialogues during the experimentation (740 with System 3 and 709 with System 4) using a combination of criteria concerning the type of interaction (call duration, task completion, system rejects, system time-outs, error recovery).

In both corpora, different scenarios for automated appointment scheduling are tested combining user initiative with system initiative. The user's or system's initiatives concern appointment proposals at different moments of the interaction progression. Common results to both corpora are presented in Subsection 5.2. Then subsections 5.2.4 and 5.3 focus respectively on the commercial and the experimental corpus.

5.1 Aims of the analysis and method

The analysis of interactions is focused on four main topics:

1. identification and description of patterns of user behaviour with the spoken dialogue system (SDS),
2. identification and description of types of errors and interactional misalignment phenomena between the user and the system,
3. description of types of interruptions and their contribution to interactional misalignment,
4. on the basis of 1), 2) and 3), suggestions of strategies for error recovery.

The analysis of the interactions in the commercial corpus had a specific aim which was to evaluate System 3 in order to formulate recommendations for future research studies.

Recommendations here are concerned with three points in particular:

1. the formulation and the segmentation of prompts;
2. grammars related to different states of the dialogue;
3. the organization of the turn-taking system – especially the places to activate speech recognition.

The analysis of the experimental corpus (subsection 5.3) aimed also to formulate recommendations for future research studies but was not limited to this dimension. It focuses in addition on the description of users' actions, especially those related to emotional sequences, learning processes and trust to technology in the interactions with the spoken dialogue system. The fact that in the experimental setting each user called the system six times makes it possible to observe if any regular changes occur in the manner that the user interacts with the system. At this point we were interested in the evolution of users' interruption practices, throughout the succession of calls, directly related to modifications in the way users participate in interaction by identifying possible transition points and by turn taking.

The specificity of our analysis, inspired by ethnomethodology [12] and Conversation Analysis (three central references on this approach [13, 14, 15] and two comprehensive presentations [16, 17]), is to take into account the entire dialogue (versus focusing on isolated question/answer sequences or on user's turn alone) and to look at how the user makes sense of the interaction, step-by-step, all through its temporal progression. This analysis focuses on practical reasoning of the user, namely on what the user orients at each moment of the dialogue's temporal unfolding in order to produce his next turn.

The method consists of exhaustive and repeated listening of all the interactions in the corpora. The aim is to identify regular interactional problems. Then we proceeded to detailed transcription - using conversation analysis conventions - and analysis of selected dialogues presenting persistent difficulties in task completion.

5.2 Common results for commercial and experimental corpora

Most of the calls result in the system setting the appointment (around 70% for the commercial corpus and more than 80% for the experimental one), sometimes involving error recovery sequences. There are very few dialogues ending in a hang up.

In the commercial setting whenever the appointment is not set after three attempts, the connection with a human operator is set up fluently from the point of view of the interaction. In order to contribute to optimize the system 3 so as to improve the interaction with the user, in section 5.2.4 we present a series of findings concerning regular incidents in the commercial corpus, followed by recommendations with a view to minimizing their occurrence or attenuating their effects. The following three subsections 5.2.1, 5.2.2 and 5.2.3 focus on common results of the analysis of interactions of both naturalistic and experimental corpora.

5.2.1 Users distrust the system speech recognition abilities

Finding: Users' natural attitude to automated dialogue technologies, as observed in 1013+ corpus, is to basically distrust the system's ability to achieve correct speech recognition.

From the user's perspective the machine works under a presumption of speech recognition error (users usually think that the machine has not understood, that "the machine is stupid"). In the face of interactional misalignments, the user makes use of his knowledge of the machine's abilities to assess its behaviour and respond adaptively. This lack of trust on the machine's speech recognition abilities frequently leads the user to repeat (sometimes with changes in prosody) or to reword his utterances instead of reconsidering the ongoing interaction.

The two transcripts below (Transcripts 5.2.1 and 5.2.2[1]) make it possible to observe how the users reason in a distrust frame in the face of two different kinds of system's behaviour. In Transcript 5.2.1 the system correctly recognises user's suggested day (line 02) and produces an appropriate answer indicating no availability. In Transcript 5.2.2, in contrast, the system fails several times to identify the correct time slot, at least from the user's point of view. In spite of these differences, both users engage in the practice of persistently rewording and repeating practices of the same temporal reference in order to set the appointment that they had initially suggested. This type of sequence gives evidence of users' distrust attitude to the system speech recognition ability and more generally to its competence to manage the engineer's calendar as a human operator would do (i.e. if they say the same time slot differently, it might work).

This is particularly clear in Transcript 5.2.1 where the system gives unambiguous feed-back of having correctly understood the suggested day (Thursday in line 03) and indicates that there are no appointments available for Thursday. Nevertheless the user keeps on insisting for an appointment on Thursday by giving more details about the date (line 05).

Time	Line	Spk	Transcription	Comments
06 :16	01	S	Essayez de spécifier d'autres contraintes quand souhaitez-vous prendre rendez-vous/	
			(3)	
	02	U	hh ahh jeudi hhh	
06 :28	03	S	vous avez une préférence pou :r ? jeudi. (.) je suis désolée (.)nous n'avons trouvé aucun créneau satisfaisant vos contraintes (.) nous allons reprendre depuis le début (.) [essayez de spécifier d'autres contraintes (.) quand souhaitez-vous prendre rendez-vous ?]	
	04	U	[hhhhhhh]	
06 :42	05	U	jeudi vingt deux hhh	

Transcript 5.2.1: User's distrust orientation to system's ASR competence (1). (1013+ EXPER : 89284- PE)

Transcript 5.2.2 below shows a slightly different situation regarding the trust issue; here the system's working is not explicit. Actually when it "hears" Friday, it suggests the first Friday available which is not always the week's Friday. For example, in line 03 the system suggests next week's Friday in response to the user's suggestion "Friday morning" (line 02, meaning Friday of this week).

The system's behaviour is based here on two problematic presumptions about the way user reasons. First of all, the assumption that if the user indicates to be available on a Friday, he will be available on every

[1]In the identification of transcribed dialogues 1013+ EXPER indicates dialogue transcripts coming from the experimental corpus; 1013+ NATUR indicates dialogue transcripts coming from the naturalistic (commercial) one. 1013+ corpus stands for all selected interactions in both corpora.

Friday is wrong. Appointment suggestions on that basis systematically result in user rejecting them in both the naturalistic and the experimental corpora. The user's pragmatic perspective regarding the organisation of his time in the week substantially differs from the system's abstract reasoning and inferences.

Second, the present design of the system ignores that temporal references in ordinary language are indexical, i.e. the mention of a day without further qualification means this day in the current week (Transcript 5.2.2, line 02) . So in those cases it is probably pointless to suggest appointments on the same day but in another week because they have a high chance to be rejected as in line 05.

These two aspects of system's design - general inferences on users' time organisation and ignoring indexicality of temporal references - constitute the origin of most interactional misalignments in the 1013+ corpus.

Time	Line	Spk	Transcription	Comments
00 :34	01	S	quand souhaitez-vous prendre rendez-vous ?	
00 :38	02	U	vendredi matin	02-U means this Friday, i.e. 16/07
00 :41	03	S	merci de confirmer la réservation du créneau suivant. le vendredi vingt trois juillet avec le passage du technicien dans l'après-midi. est-ce que cela vous [con-]	S infers that U is always available on Friday and suggests next Friday
00 :49	04	U	[non]	U rejection
00 :50	05	S	vous venez de refuser un créneau. aucun autre créneau ne satisfait vos contraintes. nous allons rependre depuis le début. essayez de spécifier d'autres contraintes. Quand souhaitez-vous prendre rendez-vous ?	
01 :05	06	U	vendredi (.) seize (.) matin	U rewords and precises line 02
	07	S	merci de confirmer la réservation du créneau suivant. le vendredi seize juillet avec le passage du technicien dans l'après-midi. (.) est-ce que cela vous convient ?	
01 :18	08	U	non	
	09	S	vous venez de refuser un créneau. aucun autre créneau ne satisfait vos contraintes. nous allons rependre depuis le début. essayez de spécifier d'autres contraintes. Quand souhaitez-vous prendre rendez-vous ?	
01 :34	10	U	vendredi (.) seize (.) matin	U repeats line 06
	11	S	vous avez une préférence pou :r vendredi de cette semaine dans la matinée. je suis désolée. (.) nous n'avons trouvé aucun créneau satisfaisant vos contraintes. (.) nous allons reprendre depuis le début. (.) essayez de spécifier d'autres contraintes. (.)quand souhaitez-vous prendre rendez-vous ?	S gets the right timeslot but indicates it's not free
	12		(4)	
01 :56	13	U	hhh° hh[hh	

Time	Line	Spk	Transcription	Comments
	14	S	[désolée. (.) je ne suis pas sûre d'avoir bien compris. (.) [...] quand souhaitez-vous prendre rendez-vous ?	
03 :11	15	U	hhh° hhhh. Hmmm (.) jeudi matin	
	16		je suis désolée(.) je ne suis pas sûre d'avoir bien compris (.) avez-vous dit ? (.) jeudi. merci de confirmer par oui ou par non.	S hears the day but not the timeslot
	17	U	Oui	
	18	S	merci de confirmer la réservation du créneau suivant. le jeudi vingt deux juillet avec le passage du technicien dans l'après-midi. (.) [est-ce que c-	
03 :33	19	U	[non	
	20	S	vous venez de refuser un créneau. aucun autre créneau ne satisfait vos contraintes. nous allons rependre depuis le début. essayez de spécifier d'autres contraintes. Quand souhaitez-vous prendre rendez-vous ?	
	21	U	jeudi matin	U repeats line 15
	22	S	vous avez une préférence pou :r jeudi. je suis désolée. (.) nous n'avons trouvé aucun créneau satisfaisant vos contraintes. (.) nous allons reprendre depuis le début. (.) essayez de spécifier d'autres contraintes. (.)quand souhaitez-vous prendre rendez-vous ?	
04 :13	23	S	lundi dix-neuf matin	

Transcript 5.2.2: User's distrust orientation to system's ASR competence (2). (1013+ EXPER : 50904- PE)

In Transcript 5.2.2 the user who asked for this Friday is supposed to be competent enough to correctly infer that this timeslot is not available while listening to system suggestion of Friday of the next week (line 03) . That is not what happens, instead, the user tries to get this timeslot by rewording and specifying twice (lines 06 and 10) the utterance he initially produced (line 02). It's only after his third formulation of the same timeslot (line 10) that he gets a clear enough answer from the system informing that this timeslot is unavailable (line 11). That finally allows the interactional partners to move on and negotiated a new appointment. The same interactional misalignment due to the system making general inferences on the basis of indexical time references used by the caller is performed in the last sequence in Transcript 5.2.2 (lines 14-22). The user questions the relevance of the system's answer as it proposes a next week slot (line 18). Then he rejects it in overlap with the system's turn (line 19). He needs to test it out again by repeating his initial suggestion (line 21, repeats line 15).

These persistent rewordings and repetitions show that users are basically suspicious about the system's utterances and that they are willing to test out their validity before changing a timeslot. That is time consuming and may be irritating for the user. This type of users' action is found in the face of interactional misalignments (as in Transcript 5.2.2) especially in the experimental corpus. It may partially be connected to the fact that in this setting, users actually try to perform a scenario with a limited number of timeslots rather than set a real appointment to get their land line problem solved.

Recommendation: Vary the system's messages grammar in order to manifest and confirm continuously

the correct recognition of user's utterance. This strategy aims to make the user feel confident about the ASR and helps her in eventually refocusing the situation and searching to locate the trouble source elsewhere (for example in Transcript 5.2.1 above there are no appointments available for Thursday so that the user has to change the day proposal instead of going on trying the same slot).

5.2.2 Conventions of temporal reference in ordinary language

Finding: The indexicality of temporal references from ordinary language is not incorporated into the system's design. By default, in ordinary language the simple mention without any other qualification of a day of the week makes reference to this day of the presently ongoing week (e.g. "Friday morning" means "this Friday in the morning").

Recommendation: To adapt the voice recognition to this convention.

Similarly to Transcript 5.2.1 and Transcript 5.2.2, in Transcript 5.2.3, the caller needs to clarify that the chosen Friday is not that of the following week, but this week's Friday. That lengthens the sequence to no use; the system must repeat, the client must reject the suggestion so repeated and finally suggest the date she thought she had conveyed at the beginning. The misalignment recalls the convention according to which the mention in the future tense of a day of the week with no other qualification makes reference to the next date that corresponds to this day on the calendar.

Time	Line	Spk	Transcription	Comments
2:35	01	S	nous avons testé votre ligne. le problème nécessite un rendez-vous avec un technicien à votre domicile. dites-nous le jour de la semaine où vous seriez présent à votre domicile, du lundi au vendredi, en précisant le matin où l'après-midi. par exemple (.) mardi matin. je vous écoute.	
	02	U	eu::h vendredi matin	i.e. next Friday, April 2 2010
	03		(1.4)	
	04	S	très bien. dans ce cas, nous vous proposons le rendez-vous suivant. le (.) vendredi (.) neuf (.) avril avec un début d'intervention du technicien entre entre huit heures? et (0.6) dix heures?= [est-ce que-]	
	05	U	=ben non [j'croyais que] c'était-	U expresses surprise
	06		(0.9)	
	07	S	je ne vous ai pas compris. afin de permettre l'intervention d'un tech[nicien sur votre ligne,]	
	08	U	[vendredi deux.]	
	09	S	nous vous proposons ce rendez-vous, le (.) vendredi (.) neuf (.) avril l'arrivée du technicien à votre domicile aura lieu entre huit heures? et (.) dix heures? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	10		(1.7)	
	11	U	non	

Time	Line	Spk	Transcription	Comments
	12		(1.2)	
	13	S	dans ce cas, dites-nous le jour de la semaine ou vous seriez présent à votre domicile, du lundi au vendredi, en [précisant le matin ou l'après-midi.]	
	14	U	[lun- vendredi deux.]	
	15		(0.5)	
	16	S	pa-	
	17		(2.2)	
	18	U	vendredi deux avril	
	19		(1.5)	
	20	S	très bien. dans ce cas, nous vous proposons le rendez-vous suivant. le (.) vendredi (.) deux (.) avril avec un début d'intervention du technicien entre treize heures? trente? et (.) quinze heures (.) trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	
4:05	21	U	oui:	
	22		(1.1)	
	23	S	très bien. nous enregistrons votre rendez-vous dans notre agenda.	

Transcript 5.2.3: Day of the week (1013+ NATUR:5)

5.2.3 Users' weighing practices and time-out tolerance

Finding: In both scenarios (system's initiative or user's initiative), the user might need time to accept or suggest a timeslot, to the extent that she might need to check her diary or look for information not immediately available (i.e. consulting another person in the household). The delay the user engages in to ponder her answer leads the system to a time-out prompt and to repeat the question.

It is interesting to note that this problem exists in both collections, i.e. the naturalistic one and the experimental one. Users might need more time to accept or propose an appointment even when as in the experimental setting they are provided with a clear calendar (see Figure 1). Compared to real life, this calendar is artificial and abstract as far as time constraints and arbitration practices regarding timeslot selection are concerned.

Regarding timeslot negotiation in the naturalistic corpus, interactional troubles might be due to off-talk (see definition below) with a person next to the caller (paragraph 5.2.3). This phenomenon is not found in the experimental corpus where interactional misalignment comes from the user's delay in pondering her answer, which might involve consulting her calendar (paragraph 5.2.3).

Naturalistic corpus: off-talk In the following dialogue (Transcript 5.2.4), after the system's suggestion the client checks with another person in her physical environment before giving an answer. The question the user addresses in off-talk to that other person (lines 04-05) is recognised as the answer to the system's question, which results in the system making a speech recognition error (line 07). Following the usual definition, we consider off-talk to be users' utterances in a dialogue that appear not to be ad-

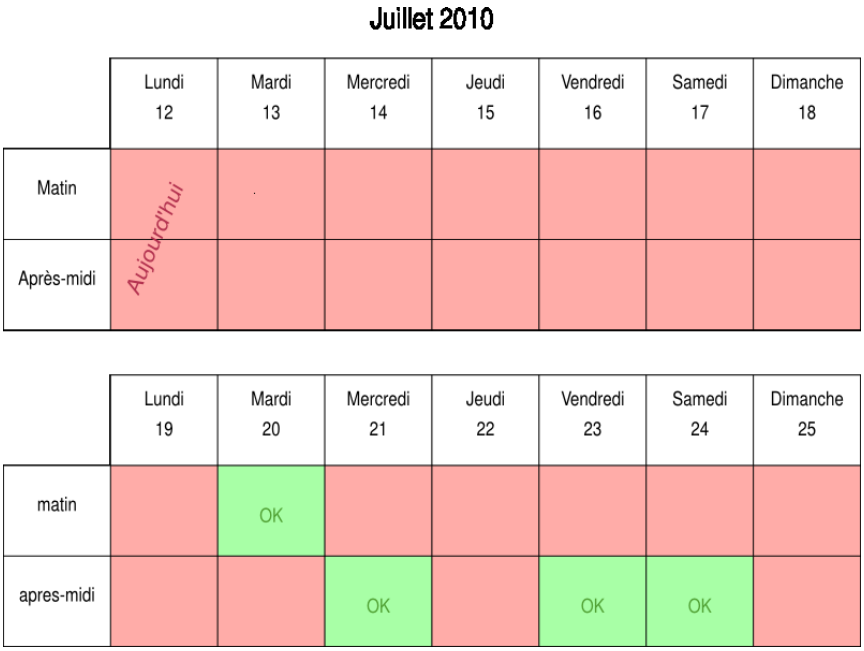

Figure 5.2.1: Example of user's personal calendar

dressed to the system. The question in off-talk is thus repeated (line 10), and the client finally accepts the suggested appointment.

Time	Line	Spk	Transcription	Comments
2:13	01	S	. . . d'après les tests que j'ai effectués sur votre ligne, l'intervention à votre domicile d'un technicien est nécessaire. j'ai un premier rendez-vous disponible demain (.) jeudi (.) premier (.) avril avec un début d'intervention du technicien entre neuf heures(.) trente? e:t (.) onze heures? trente? est-ce que cela vous convient?	
	02	U	e- [est-ce qu'il] peut venir demain matin&	
	03	S	[merci de me-]	
	04	U	&le technicien	U checks with somebody else.
	05	U'	bien sû:r	
	06	U	oui [pas de problème.]	U accepts S's suggestion
	07	S	[je ne vous ai pas com]pris	recognition error
	08		(0.5)	
	09	U	[oui:]	

Time	Line	Spk	Transcription	Comments
	10	S	[afin] de permettre l'intervention d'un technicien sur votre ligne, je vous propose ce rendez-vous, demain (.) jeudi (.) premier (.) avril avec un début d'intervention du technicien entre neuf heures (.) trente? e:t (.) onze heures? trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	11		(0.8)	
3:01	12	U	oui:	
	13		(0.3)	
	14	S	très bien. j'enregistre votre rendez-vous dans notre agenda.	

Transcript 5.2.4: Off-talk (1013+ NATUR:3)

The following dialogue (Transcript 5.2.5) allows observing again how off-talk creates trouble in the interaction. The caller checks with another person before accepting the appointment suggestion. Her off-talks are treated as the answer that the system expects, which gives rise to a recognition failure that forces the system to repeat the prompt (line 04).

Time	Line	Spk	Transcription	Comments
2:14	01	S	d'après les tests que nous avons effectués sur votre ligne, l'intervention à votre domicile d'un technicien est nécessaire. nous avons un premier rendez-vous disponible, demain (.) jeudi (.) premier (.) avril avec un début d'intervention du technicien entre quinze heures (.) trente? e:t (.) dix-sept heures (.) trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	02		(2.0)	
	03	U	°entre: quinze heures trente e:::t dix-sept heures°=	U checks with somebody else
	04	S	=je ne vous ai pas compris. afin de permettre l'intervention d'un technicien sur votre ligne, nous vous proposons ce rendez-vous, demain (.) jeudi (.) premier (.) avril l'arrivée du technicien à votre domicile aura lieu entre quinze heures (.) trente? e:t (.) dix-sept heures (.) trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	recognition error
	05	U	non (XX XX XX) c'est une boîte	
	06		(1.2)	
	07	S	je ne vous ai pas compris. pour [votre] rendez-vous avec un&	
	08	U	[oui]	
3:12	09	S	&technicien, nous vous remercions de patienter, notre conseiller va vous répondre.	

Transcript 5.2.5: Off-talk 2 (1013+ NATUR:4)

Experimental corpus: pondering delay Even in the experimental corpus, with its clear, abstract and fictitious calendar (see Figure 5.2.1), it is not always easy for the user to suggest an appointment in time. In timeslot negotiation he often takes time to consider different possibilities. Meantime system treats these pondering delays as time-out and takes the floor (Transcript 5.2.6: lines 03-04, 07-08, 14-15; Transcript 5.2.7: lines 03-04). That leads to waste of time as system has to repeat the initial question (as in Transcript 5.2.6 line 04 repeats question in line 01).

Time	Line	Spk	Transcription	Comments
3:30	01	S	je suis désolée (.) nous n'avons trouvé aucun créneau satisfaisant vos contraintes. nous allons reprendre depuis le début. essayez de spécifier d'autres contraintes. quand souhaitez-vous prendre rendez-vous ?	
	02		(3)	
3:43	03	U	°eeu ::h j'ai plus de créneaux moi°=	pondering delay put in words
3:46	04	S	=désolée (.) je ne suis pas sûre d'avoir bien compris	
	05		(5)	pondering delay
3:52	06	U	hh° hh [jeudi :: vingt deux ap-	
	07	S	[désolée (.) je ne suis pas sûre d'avoir bien compris. quand souhaitez-vous prendre rendez-vous ?=	
4:00	08	U	=°j sais plus°	pondering delay put in words
	09		(5)	
	10	S	désolée (.) je ne suis pas sûre d'avoir bien compris (.) procédons différemment (.) je vous propose le mercredi quatorze juillet avec le passage du technicien dans l'après-midi.	
	11		. . .	
4:23	12	S	quand souhaitez-vous prendre rendez-vous ?	
	13		(6)	
4:29	14	U	[le sam-	
	15	S	[désolée (.) je ne suis pas sûre d'avoir bien compris (.) procédons différemment (.) je vous propose. . .	

Transcript 5.2.6: Pondering delay 1 (1013+ EXPER: 34374-PE)

Time	Line	Spk	Transcription	Comments
1:52	01	S	quand souhaitez-vous prendre rendez-vous/	
	02		(5)	
1:57	03	U	hhh hmm[m ::	U ponders a date
1:58	04	S	[désolée (.) je ne suis pas sûre d'avoir bien compris (.) je répète quand souhaitez-vous prendre rendez-vous/	S takes the floor too early
	05	U	hh jeudi quinze après-midi	

Time	Line	Spk	Transcription	Comments
	06	S	vous avez une préférence pour jeudi pour cette semaine dans l'après-midi	

Transcript 5.2.7: Pondering delay 2 (1013+ EXPER: 50904-PE)

Pondering delay is manifest in relatively long silences before user's taking his turn, then results in latched or overlapped turns (Transcript 5.2.6: lines 03-04, 07-08, 14-15; Transcript 5.2.7: lines 03-04). In Transcript 5.2.6 user expresses verbally the reasons of this delay in lines 03 and 08 and the fact that she is searching for timeslots in her calendar, which is why she cannot answer immediately.

Recommendation: Lengthen the user's slot after presenting the suggestion (i.e. relax the delay before the time-out).

5.2.4 Commercial test of system 3: results and recommendations

Users' difficulties in retaining appointment details

Finding: The system's suggestions might turn out to be hard to retain.

In the sequence below (Transcript 5.2.8) misalignments arise due to the difficulty that the caller experiences to retain the various details that make up the suggestion "tomorrow Thursday April 1st, beginning of the intervention between 2 pm and 4 pm"). The caller fails to understand twice (lines 03 and 10). The system initiates recovery after the first incomprehension (lines 05-09), but mistakenly takes the second as an acceptance (line 12). The system ends up registering the suggested appointment in spite of the fact that the caller still wonders which was the accepted date (line 18).

Time	Line	Spk	Transcription	Comments
1:42	01	S	d'après les tests que nous avons effectués sur votre ligne, l'intervention à votre domicile d'un technicien est nécessaire. nous avons un premier rendez-vous disponible, demain (.) jeudi (.) premier (.) avril avec un début d'intervention du technicien entre quatorze heures (.) e:t (.) seize heures est-ce que cela vous convient? merci de répondre par oui ou par non.	
	02		(1.9)	
	03	U	eu:h je n'ai pas compris.	U does not understand
	04		(0.8)	
	05	S	je ne vous ai pas compris. (0.5) a[fin de per]mettre l'intervention d'un&	
	06	U	[je: non-]	
	07	S	&technicien sur votre ligne, (0.25) nous [vous propos]ons ce rendez-vous,	
	08	U	[oui oui mais-]	

Time	Line	Spk	Transcription	Comments
	09	S	demain (.) jeudi (.) premier (.) avril l'arrivée du technicien à votre domicile aura lieu entre quatorze heures (.) e:t (.) seize heures est-ce que cela vous convient? merci de répondre par oui ou par [non.]	
	10	U	[mais-]	
	11		(0.7)	
	12	S	très bien. (0.4) pen[dant que nous enregistrons votre&	
	13	U	[cela me convenait v:: &	
	14	S	& rendez]-vous dans notre agenda,	
	15	U	& quand?]	
	16	S	merci de patienter quelques instants	
	17		(2.1)	
2:39	18	U	c'est pas possib' on peut pas avoir un::. quelqu'un au bout du fi:l?	

Transcript 5.2.8: Retention troubles (1013+ NATUR:1)

In the following dialogue (Transcript 5.2.9) the system succeeds in transferring the call to a live agent timely, but again misalignments arise because of the caller's inability to retain all the details that regard the suggested appointment. This time, the caller is not able to properly grasp the starting time of the technician's intervention; that happens twice (lines 02, 07 and 09).

Time	Line	Spk	Transcription	Comments
1:54	01	S	d'après les tests que nous avons effectués sur votre ligne, l'intervention à votre domicile d'un technicien est nécessaire. nous avons un premier rendez-vous disponible, le (.) vendredi (.) deux (.) avril avec un début d'intervention du technicien entre huit heures? trente (.) e:t (.) dix heures? trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	02	U	excusez-moi, vous m'avez dit huit heures trente neuf heures trente ou [dix heures trente]	U is not certain about her understanding of the system's prompt
	03	S	[je ne vous ai pas] compris.	
	04		(0.5)	
	05	U	[oui]	
	06	S	[a]fin de permettre l'intervention d'un technicien sur votre ligne, nous vous proposons ce rendez-vous, le (.) vendredi (.) deux (.) avril avec un début de l'intervention du technicien entre huit heures? trente? e:t (.) dix heures? trente? est-ce que cela vous convient? merci de répondre par oui ou [par non.	
	07	U	[alors excusez]-moi:, (0.3) je vous fais répéter une deuxième fois,=	
	08	S	=je ne [vous ai pas com]pris.	

Time	Line	Spk	Transcription	Comments
	09	U	[je n'ai pas] je n'ai pas enten[du si vous me disiez-]	
2:50	10	S	[pour votre rendez-vous avec un] technicien, nous vous remercions de patienter, notre conseiller va vous répondre.	

Transcript 5.2.9: Retention troubles 2 (1013+ NATUR:2)

Recommendation: Given the possibility of not being able to retain all the details of the suggested appointment, add a further option to yes/no to make it possible to hear those details again. For example, add the option "say repeat" to the yes/no alternative.

5.2.5 Users' practice of repeating after reject notification

Finding: In the face of recognition errors regarding yes/no, callers tend to treat the silence that follows the system's turn "I didn't understand" as a possible transition place (i.e. as a possible permutation point between hearer and listener). They avail of this silence to take the floor and repeat their choice (i.e. yes or no). They sometimes amount to rewordings (e.g. Transcript 5.2.4: "of course" in line 05 becomes "yes" in line 06). However, at that location the system "does not hear" (voice recognition is disabled), which results in the system missing an opportunity for quick error recovery.

See above Transcript 5.2.4 (lines 05-08) and Transcript 5.2.5 (line 08); below Transcript 5.2.14 (line 17).

Recommendation: To lengthen the silence after "I didn't understand" while activating voice recognition of expressions of acceptance or refusal.

5.2.6 The logical relationship between turns at talk

Finding: Users have to understand the link between their suggestion and the system's counter-suggestion. Now, the system's phrase "In this case…" opens up the possibility of an illusory logical relationship between the user's and the machine's respective suggestions.

In the dialogue below (Transcript 5.2.10), the caller chooses Friday morning (line 03) in response to the system's prompt to select a day and a time for the technician's intervention. She gets in response an appointment suggestion for Monday morning (line 05). The obvious lack of relationship between a suggestion for Friday (the client's) and a counter-proposition for Monday (the system's) is aggravated by the connector "In that case", which seems to presuppose a logical connection between the two terms . Presumably due to this difficulty of understanding the caller does not respond after the presentation of the counter-suggestion (see silence in line 07). The following two suggestions by the system are refused (lines 10 and 14), and the call ends up being transferred to a live agent (line 16).

Time	Line	Spk	Transcription	Comments
2:35	01	S	nous avons testé votre ligne. le problème nécessite un rendez-vous avec un technicien à votre domicile. dites-nous le jour de la semaine où vous seriez présent à votre domicile, du lundi au vendredi, en précisant le matin ou l'après-midi. par exemple (.) mardi matin. je vous écoute,	

Time	Line	Spk	Transcription	Comments
	02		(1.9)	
	03	U	°(xx xx) vendredi?° (0.7) vendredi matin.	
	05		(1.6)	
	06	S	très bien. dans ce cas, nous vous proposons le rendez-vous suivant. le (.) lundi (.) dix-neuf (.) avril avec un début d'intervention du technicien entre huit heures? e:t (.) dix heures? est-ce que cela vous convient? merci de répondre par oui ou par non.	Here, "dans ce cas" prefaces a suggestion that holds no obvious link with U's suggestion in line 03
	07		(2.7)	
	08	S	je ne vous ai pas entendu. afin de permettre l'intervention d'un technicien sur votre ligne, nous vous proposons ce rendez-vous, le (.) lundi (.) dix-neuf (.) avril avec un début d'intervention du technicien entre huit heures? e:t (.) dix heures? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	09		(0.7)	
	10	U	non	
	11		(1.4)	
	12	S	dans ce cas, nous vous proposons un autre rendez-vous possible, le (.) mardi (.) treize (.) avril l'arrivée du technicien à votre domicile aura lieu entre quatorze heures? e:t (.) seize heures? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	13		(0.9)	
	14	U	non	
	15		(1.2)	
3:48	16	S	pour votre rendez-vous avec un technicien, nous vous remercions de patienter, notre conseiller va vous répondre.	

Transcript 5.2.10: Unclear logical links 1 (1013+ NATUR:7)

Similarly, in the dialogue below (Transcript 5.2.11) the caller's suggestion for Monday morning (line 03) is followed by a counter-suggestion for Wednesday morning (line 05). The caller responds with irritation (line 06); the system is not able to process the response (line 08). Then the system repeats the counter-suggestion (08), which the caller rejects (line 10). The call is finally transferred to a live agent.

Time	Line	Spk	Transcription	Comments
1:45	01	S	d'après les tests que j'ai effectués sur votre ligne, l'intervention à votre domicile d'un technicien est nécessaire. merci de me préciser un jour de la semaine où vous seriez présent à votre domicile, du lund au vendredi, en précisant le matin ou l'après-midi. par exemple (.) mardi matin. c'est à vous.	
	02		(0.7)	

Time	Line	Spk	Transcription	Comments
	03	U	hh° m: lundi matin	
	04		(2.0)	
	05	S	très bien. dans ce cas, je vous propose le rendez-vous suivant. le (.) mercredi (.) seize (.) décembre avec un début d'intervention du technicien entre huit heures? et (.) dix heures? [est-ce que cette pro-	Here, "dans ce cas" prefaces a suggestion not obviously linked to U's suggestion in line 03.
	06	U	[oui mais c'est pas possib' ça hein]	
	07			(0.9)
	08	S	je ne vous ai pas compris. afin de permettre l'intervention d'un technicien sur votre ligne, je vous propose ce rendez-vous, le (.) mercredi (.) seize (.) décembre avec un début d'intervention du technicien entre huit heures? et (.) dix heures? est-ce que cela vous convient? merci de me répondre par oui ou par non.	
	09		(1.0)	
	10	U	non	
	11		(1.4)	
	12	S	dans ce cas, merci de me préciser un jour de la semaine où vous seriez présent à votre domicile, [(0.3) du] lundi au vendredi,	
	13	U	[( )]	
	14	S	dans ce cas, je vous propose le rendez-vous suivant. demain (.) vendredi (.) onze (.) décembre avec un début d'intervention du technicien entre dix heures? trente? et (.) douze heures (.) trente? est-ce que cela vous convient? merci de me répondre par oui ou par non.	Here, "dans ce cas" prefaces a suggestion that responds to a noise in line 13.
	15	U	pf::: hh° eu::h non début d'après-midi.	
	16		(0.9)	
3:18	17	S	je ne vous ai pas compris. pour votre rendez-vous avec un technicien, je vous remercie de patienter, notre conseiller va vous répondre.	
	18			
	19			
	20			
	21			
	22			
	23			
	24			
	25			

Transcript 5.2.11: Unclear logical links 2 (1013+ NATUR:8)

Recommendation: Precede "In that case…" of a feedback notification. For example: "that day of the week is no longer available" or "that first available appointment is on [day] [time]". If possible, activate

the voice recognition at the end of this message. If the notification communicates information with no apparent relationship with the user's utterance, there is a chance that the user will contribute a rejection as she perceives the first next silence.

5.2.7 Urgency as a reason for rejecting an appointment

Finding: Users sometimes reject system's suggestions by specifying that their problem requires urgent treatment; to that purpose they make use of the semantic field of urgency, e.g. "it's late", "too late", "I need my phone immediately", "it's urgent".

In Transcript 5.2.12, even thought the caller has heard the instruction to answer by "yes" or "no", she cannot help expressing her reject (i.e. a special way of saying "no") through an explanation conveying a sense of urgency (line 06). Not only does the system fail to recognise the answer (line 07), but also the caller's upset gets aggravated as the system repeats the same suggestion (line 09). The call is finally transferred to a live agent.

Time	Line	Spk	Transcription	Comments
1:57	01	S	j'ai testé votre ligne. le problème nécessite un rendez-vous avec un technicien à votre domicile, je vous propose, le (.) samedi (.) trois (.) avril avec un début d'intervention entre huit heures? et (.) dix heures? est-ce que cette proposition vous convient? merci de me répondre par oui ou par non.	
	02		(0.7)	
	03	U	non	
	04		(1.3)	
	05	S	dans ce cas, je vous propose, le (.) samedi (.) trois (.) avril avec un début d'intervention entre huit heures? trente? et (.) dix heures? trente? est-ce que cette proposition vous convient? merci de me répondre par oui ou par non.	
	06	U	trop ta:rd. j'ai besoin de m:on téléphone immédiatement.	U rejects the suggestion by explaining the reason why she rejects it.
	07	S	je n' vous ai pas compris. (0.55) afin de [per]mettre l'intervention d'un&	
	08	U	[n-]	
	09	S	&technicien sur votre ligne, je vous propose ce rendez-vous, le (.) samedi (.) trois (.) avril avec un début d'intervention du technicien entre huit heures? trente? et (.) dix heures? trente? est-ce que cela vous convient? (0.4) mer[ci de ré]pondre-	
	10	U	[no:n]	
	11		(0.7)	
3:02	12	S	pour votre rendez-vous avec un technicien, je vous remercie de patienter, notre conseiller va vous répondre.	

Transcript 5.2.12: Urgency 1 (1013+ NATUR:9)

Similarly, in Transcript 5.2.13 the caller does not limit herself to say “no”, but does so by expressing the reason of her rejection. She takes the floor three times (lines 03, 06 and 09) to reject two system’s suggestions. The call must finally be transferred to a live agent.

Time	Line	Spk	Transcription	Comments
2:12	01	S	j’ai testé votre ligne. le problème nécessite un rendez-vous avec un technicien à votre domicile, je vous propose, le (.) samedi (.) dix (.) avril avec un début d’intervention entre huit heures? et (.) dix heures? est-ce que cette proposition vous convient? merci de me répondre par oui ou par non.	
	02		(1.9)	
	03	U	samedi dix c’e:st c’est ta:rd pf::	U rejects and explains her rejection.
	04		(0.8)	
	05	S	je ne vous ai pas compris afin de permettre l’intervention d’un technicien sur votre ligne, je vous propose ce rendez-vous, le (.) samedi (.) dix (.) avril avec un début d’intervention du technicien entre huit heures? et (.) dix heures? est-ce que cette proposition vous convient? [(0.5)] Mer	
	06	U	[non] (0.9) non [c’est trop tard]	idem
	07		(0.9)	
	08	S	[dans ce cas] je vous propose, le samedi (.) dix (.) avril avec un début d’intervention entre huit heures? trente? et (.) dix heures? trente? est-ce que cette proposition vous convient? merci de répondre par oui ou par non.	
	09	U	non	
	10		(1.4)	
	11	S	pour votre rendez-vous avec un technicien, je vous remercie de patienter, notre conseiller va vous répondre.	
3:18	12	U	merci	

Transcript 5.2.13: Urgency 2 (1013+ NATUR:10)

Recommendation: To try to take into account some vocabularies of motives so as to manage the urgency of the client’s request. To incorporate in the grammar of the voice recognition some expressions from the semantic field of urgency (“it’s urgent”, “it’s too late”, “immediately”, . . .) in order to give the initiative to the client whenever she evaluates in these terms the first system’s suggestion.

5.2.8 Users’ practice of transforming examples into suggestions

Finding: Users might understand an example as a suggestion or as a hypothesis that they have to accept or reject.

In the sequence below (Transcript 5.2.14), the caller, having rejected the system’s first suggestion, is invited to suggest herself a date and a time. Instead she takes the example (Tuesday morning) as a new

suggestion that she ends up rejecting while explaining the reason of the reject (line 07). The system must notify a recognition problem and repeat the instruction.

Time	Line	Spk	Transcription	Comments
1:53	01	S	d'après les tests effectués sur votre ligne, un rendez-vous avec un technicien est nécessaire à votre domicile. le premier rendez-vous disponible est le suivant, demain (.) jeudi (.) premier (.) avril (.) le début d'invervention du technicien aura lieu entre treize heures (.) trente? et (.) quinze heures (.) trente? est-ce que cela vous convient? merci de répondre par oui ou par non.	
	02		(2.1)	
	03	U	non	
	04		(1.3)	
	05	S	dans ce cas, merci de me dire un jour de la semaine où vous seriez présent à votre domicile, du lundi au vendredi, en précisant le matin ou l'après-midi. par exemple (.) mardi matin. c'est à vous.	
	06		(2.5)	
	07	U	avant	U declines the example, which is taken as a new suggestion (i.e. Tuesday April 6 instead of Thursday April 1)
	08		(1.9)	
	09	S	je ne vous ai pas entendu. afin de permettre l'intervention d'un technicien sur votre ligne, merci de me dire un jour de la semaine où vous seriez présent à votre domicile, de lundi à vendredi, en précisant le matin ou l'après-midi. par exemple (.) mardi matin. c'est à vous.	S is not able to process "avant" as a response to its prior turn.
	10		(1.6)	
	11	U	ven<u></u>dredi matin.	the emphasis on "vendredi" suggests a contrast with "mardi".
	12		(1.5)	
	13		très bien. dans ce cas, je vous propose le rendez-vous suivant. (0.6) le (.) vendredi (.) deux (.) avril (.) avec un début de l'intervention du technicien, entre huit heures? (.) et (.) dix heures? est-ce que cela vous convient?	

Time	Line	Spk	Transcription	Comments
	14	U	très bien	U anticipates the end of S's turn, therefore U is not invited to express her acceptance by "yes". In its stead, U imitates the system's way of accepting. (cf. line 13)
	15		(1.0)	
	16	S	je ne vous ai pas compris. (0.6) afin [de perme]ttre l'intervention d'un&	S notifies processing difficulties
	17	U	[oui:]	U rewords her acceptance, taking the silence in line 16 as transition relevance place.
	18	S	&technicien sur votre ligne, un rendez-vous est nécessaire. (0.6) merci de me dire si vous êtes disponible, (0.5) le (.) vendredi (.) deux (.) avril (.) avec un début de l'intervention du technicien, entre huit heures? (.) et (.) dix heures? oui ou non?	
	19		(1.0)	
	20	U	oui	
	21		(1.1)	
3:32	22	S	très bien. nous enregistrons votre rendez-vous dans notre agenda.	

Transcript 5.2.14: Example as suggestion (1013+ NATUR:12)

Recommendation: Include in the grammar of the relevant dialogue state expressions used to accept suggestions, then associate these words to the concept embodied by the example. In other words, the grammar for the prompt "Please tell me a day of the week…", given that the prompt finishes with the example "Tuesday morning", could include expressions like "okay", "that's fine", …as associated to the concept Day[Tuesday] Time[morning].

5.2.9 Users' manifold ways of accepting an appointment

Finding: Users might express acceptance of an appointment in very different ways (e.g. "yes", "okay", "fine", …). Furthermore, the system uses the formula "very well" for confirmation, which invites imitation practices on the part of the human user that have been largely documented. However, the system only accepts "yes" as a valid confirmation for a suggestion and notifies error if confirmation takes on any other shape (e.g. in Transcript 5.2.14 above, line 14). As a result, the system repeats its suggestion (Transcript 5.2.14, lines 16-18), which lengthens the interaction.

Recommendation: To enrich the semantic field of recognition of appointment acceptances to fit linguistic usages.

5.3 Final experimental test of Systems 3 and 4: results and recommendations

5.3.1 Learning how to talk to machines

The examination of collections of six calls to the system made by each user in the experimental setting allows focusing on some regular changes in users' methods of interaction. In most of the cases, the analysis of these successions of six dialogues makes it quite clear that the user progressively learns to interact with the system, to the extent that she appears to acquire knowledge and familiarity. This can be observed in particular from modifications in her ways to take the floor in the interaction and to interrupt the system's turns at precise points, which we designate possible transition relevance places. Relevant interruptions are possible by virtue of the knowledge, acquired through experience, that the system's turns may be overlapped without compromising the interaction; they are also made possible by the ability the user acquires to anticipate the complete system's prompts. Interruption practices occur after the first or second call and then become stabilized for the rest of dialogues. Competent interruptions make it possible to save time while progressing in task achievement.

As a rule, users learn to interrupt the system's welcome message (Transcript 5.3.1) and yes/no questions (Transcript 5.3.2).

Transcript 5.3.1 contains the openings of five successive calls of the same user and the sequential positioning of his first action: typing the identification code. In the first two calls user listens to the entire message and awaits the system's instruction to type the code. Then, starting from the third call he systematically interrupts the system's welcome message around the same sequential position. This is a typical behaviour observed in the corpus; some users interrupt welcome message from the second call onwards.

Time	Line	Spk	Transcription	Comments
			1st call	
	01	S	bonjour. bienvenue sur l'expérimentation européenne CLASSIC . veuillez taper le code de votre appel suivi de la touche étoile.	
	02	U	((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	TUT: Stands for 1 digit DTMF dialing
	03	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	
			2nd call	
	04	S	bonjour.bienvenue sur l'expérimentation européenne CLASSIC . veuillez taper le code de votre appel suivi de la touche étoile.	
	05	U	((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	
	06	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	

Time	Line	Spk	Transcription	Comments
			3rd call	
	07	S	bonjour.bienvenue sur l'expérimentation européenne [CLASSIC .	
	08	U	[((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	Overlapping and interruption
	09	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	
			4th call	
	10	S	bonjour.bienvenue sur l'expérimentation européenne [CLASSIC .	
	11	U	[((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	Overlapping and interruption
	12	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	
			5th call	
	13	S	bonjour.bienvenue sur l'expérimentation européenne CLASSIC . [veuillez	
	14	U	[((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	Overlapping and interruption
	15	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	

Transcript 5.3.1: Interrupting system's welcome message (1013+ EXPER, D)

Transcript 5.3.2 below shows excerpts of the fourth call of the user presenting the regular practice of overlapping concerning system's yes/non questions: lines 05, 08 and 11.

Time	Line	Spk	Transcription	Comments
0:00	01	S	bonjour.bienvenue sur l'expérimentation européenne CLASSIC . [veuillez	
	02	U	[((TUT TUT TUT TUT TUT TUT TUT TUT TUT))	Overlapping and interruption
	03	S	vous allez être redirigé vers le service de prise de rendez-vous. (.)merci de patienter.	
			[. . .]	
0:24	04	S	le premier rendez-vous disponible est le mercredi quatorze juillet dans l'après-midi. est-ce que cela vous convient ? [merci de repond-	
	05	U	[non	Overlappling Yes/No question

Time	Line	Spk	Transcription	Comments
	06	S	dans ce cas merci de me dire un jour de la semaine ou vous seriez présent à votre domicile	
			[. . .]	
0:53	07	S	bon je vous propose le vendredi seize juillet dans l'après-midi. est-ce que cela vous convient ? [merci	
	08	U	[non	Overlappling Yes/No question
	09	S	dans ce cas merci de me dire un jour de la semaine ou vous seriez présent à votre domicile	
			[. . .]	
1:16	10		très bien. je vous propose donc le mercredi vingt et un juin dans la matinée. est-ce que cela vous convient ? [merci d-	
	11	U	[oui	Overlappling Yes/No question
	12	S	très bien (.) j'enregistre votre demande dans mon agenda	

Transcript 5.3.2: Overlapping the yes/no question (1013+ EXPER: 78858 - PE)

During his first call this user does not interrupt the system's turns. Starting from his second call he always interrupts the welcome message and proceeds to different overlaps during the dialogues, namely of yes/no questions as in Transcript 16. An important observation is that overlapping is produced at the point in time of the system's message immediately following the question: "est-ce que cela vous convient ?" (lines 04, 07, 10). This demonstrates that overlapping is not a random behaviour but a regular action structured by user's orientation to possible transition relevance places in the machine's talk. Competence to correctly identify these places is contingent upon experience in using the SDS.

5.4 Playful error recovery

Finding: When they respond emotionally to unexpected system responses, some users spontaneously move from irritation to laughter.

Recommendation: Design system prompts able to encourage playful error recovery.

Combining CA methods [12, 13, 14, 15, 16, 17] with ATE notions [18, 19] we have attempted to show that emotional responses in user-system dialogues vary according to the degree of control they display and to the nature of the transformations they undergo.

Of special interest in the present context is the contrast between emotional transformations that involve the transition from irritation to resignation and emotional transformations that regard the shift from irritation to laughter.

In our previous corpus studies we have been able to identify two kinds of transformed emotional response, which differ drastically as far as their interactional import is concerned. The first transformation regards responses that go from anger to resignation. Characteristic of angry behaviours is a sense of removing an obstacle that impedes goal achievement; resignation, in contrast, implies a sense of loss, like any other form of sadness, but also the appraisal that nothing can be done, in our case, that the obstacle cannot

actually be removed. When users that initially get upset because of unexpected responses by the system move towards resignation, they also move towards a strong motivation to quit the dialogue. There are several forms resignation can take on. For example, we have documented a number of instances in which the user, faced with unexpected system responses, engages in a proverbial form of angry resignation that finally motivates him to hang up: overtly insulting the system.

Let us look now at an emotional transformation from the 1013+ experiment that takes place in an analogous situation but ends up affecting the course of the dialogue in a radically different fashion.

Time	Line	Spk	Transcription	Comments
	01	S	quand souhaitez-vous prendre rendez-vous?	
	02		(0.9)	
	03	U	Samedi dix-sept juillet après-midi h	
	04		(1.6)	
	05	S	désolé:, je ne suis pas sûr d'avoir bien compr[is. (0.6)]	
	06	U	[°ah putain°]	
	07	S	avez-vous dit, samedi de cette semaine dans l'après-midi. merci de confirmer par oui ou par non.	
	08		(0.8)	
	09	U	oui	
	10		(3.1)	
	11	U	oui	
	12	S	je ne vous ai pas entendu. je répète. [(0.7) avez-][vous dit,]&	
	13	U	[h::] [((laugh))]	
	14	S	& samedi de cette semaine dans l'après-midi. merci de confirmer par oui ou par non.	
	15		(0.6)	
	16	U	oui:	
	17		(1.6)	
	18	U	[oui:]	
	19	S	[vous avez] une préférence pour samedi de cette semaine dans l'après-midi.	

Transcript 5.4.1: From irritation to laughter (1013+ EXPER: 98206)

Transcript 5.4.1 is analogous to dialogues we have previously studied insofar as it involves a system that provides unexpected responses (including repetitions) to a user who thinks he is doing the right thing, and so responds emotionally to what he senses to be the system's inability to make the dialogue progress. The first emotional act occurs in line 06 as the system notifies difficulty in processing the user's last input and asks for confirmation: the user off-talks a swearword, in sign of irritation. He then responds by accepting in line 09, but as the system takes more that three seconds to get back, the user repeats the acceptance in line 11. In line 12 the system responds as though a time-out had taken place, which ostensibly further upsets the user: he responds with an irritated breathe-out in line 13. But instead of staying in that upset emotion, which he appears to have done from line 07 to line 13, he suddenly shifts to laughter in the same line. The user's next response is not emotionally marked: he accepts again the system's request for

confirmation in line 15; similarly, as the system does not respond after one and a half seconds he repeats his acceptance in line 17, which happens to occur in overlap with the system's positive response in line 19.

In other words, this transcript confronts a user to a situation where she is upset because of an apparently unjustified unexpected response by the system. But the different profiles in emotional transformation cause the emotional response to bring about radically different interactional effects. In the first case, irritation becomes angry resignation, which ends up with the user hanging up. In the above example, irritation becomes laughter, which ends up with the recovery of the error and the dialogue moving on to the next step. In this sense, the irritation-to-resignation transformation disengages the user from the dialogue, whereas the irritation-to-laughter transformation renews the threatened engagement in the dialogue.

The crucial fact is that while the origin of emotional transformation is quite predictable, i.e. irritation, the target of the transformation, i.e. resignation or laughter, cannot be told in advance. However, there might be factors that encourage either of the targets in emotional transformation. An interesting observation is that in the experimental database the chance of finding instances of emotional responses involving transformations from irritation to laughter is substantially higher than in the real-life database. Conversely, the chance of finding instances of emotional responses that shift from irritation to resignation, especially angry resignation, is higher in the real-life database than in the experimental one.

This might be taken to suggest that users that participate in experiments are more easily inclined to adopt a "play" attitude, since nothing is really at stake. Because emotions are intrinsically related to concerns, it might not make much sense to these participants to take too seriously their experimental task if that implies engaging in negative emotional experiences. The shift to laughter when irritation arises is a way of making the situation more emotionally tolerable. On the other hand, users can easily operate the shift by reminding themselves that the dialogue is "just" part of an experiment, and that nothing in their real lives will really change whether the result is one or the other.

However, reminding oneself that nothing is really at stake is not the only way in which the irritation-to-laughter transformation can be triggered. Instances of real-life dialogues show that users may operate the shift even when the situation is serious. Understandably, in the latter cases the transformation is harder; but it is not impossible. While in experimental settings a given probability of moving from irritation to laughter can be taken for granted, in real-life settings achieving the same probability might require active intervention on the part of the designer. As we saw, the advantages are at least of two kinds: first, the positive advantage of putting users in a mood that is conducive to error recovery and overall success; second, the negative advantage of preventing users from moving towards the disengaging emotion of resignation.

In order to increase the chance of having users shift from irritation to laughter in real-life settings, error notification messages can be designed to convey a sense of play to the user. Of course, it would be ridiculous to suggest that the system should at any time of the dialogue present itself as playful; this would be obviously counter-productive, to the extent that users do take the dialogue seriously and consequently expect the system to manifest the same attitude. But it could be worthwhile to program the system to contribute playful messages in the face of acknowledged errors, especially if the error is not the first one in a raw. Our hunch is that these playful messages will encourage the user to shift to laughter, which as we saw is conducive to error recovery and overall success.

The sequential analysis of emotional responses suggests that designers can make a difference by programming systems to encourage playful error recovery. But, admittedly, tailoring prompts with the right combination of instrumental sincerity and relaxing humour is not a trivial task.

5.4.1 The problems of encouraging complex utterances

Finding: The system encourages complex utterances but ignores most of the information surrounding the day formulation, which leads it to incorrect appointment proposals. As a consequence, user's rejections are observed along with elaboration of misalignment management methods.

Often, this implies a significant increase in the call duration and irritation or other kind of emotional behaviour of the user related to the repetitiveness of system's messages and the impression not to be heard.

Actually, the system's inquiry of a convenient appointment starts with an open question ("When would you like to set an appointment?") which encourages users to talk without any constraint and to give different details (for example for the same day different formulations can be found: "Thursday", "next Thursday", "Thursday next week", "Thursday 22nd of July", "the 22nd of July". Yet, in most cases, the system appears to "hear" only the day and to ignore the rest. Then, it produces an answer manifesting an interactional misalignment as it orients to a different day than the user asked for (see Transcript 5.4.2 below, lines 02-05).

It is worth noting that this interactional misalignment is far more present in the experimental corpus than in the commercial one. That is no doubt related to the fact that in the experimental corpus at about half of the dialogues are realised with system 4 which uses more open interaction strategies then the system 3. As a consequence, users' answers to system 4 are less standardized compared to system 3, which increases the risk of ASR errors.

Two main users' strategies are observed in order to manage this interactional misalignment and to move on in appointment scheduling process: switching the day (next paragraph) and rewording the temporal reference to the same day (the paragraph after).

Managing interactional misalignment: switching the day The excerpt presented in Transcript 5.4.2 below comes from a dialogue that lasted more than five minutes - an extremely long dialogue, since one minute should suffice to set the appointment. It is composed of four sequences involving negotiation of four different days and timeslots for possible appointment (respectively lines 02-05; 07-09; 11-13; 15-17). It begins with two identically structured sequences, each one culminating in the user's rejection of the appointment (lines 07-09; lines 11-13). Each of these sequences begins with the system's timeslot question, followed by the users' suggestion of a specific timeslot (respectively lines 07, 11). Then the system systematically misunderstands the day suggested by treating only the day reference and ignoring the date-related details in the user's turn. So the system suggests appointments the same day but the next week (lines 08 and 12) followed by the user's rejections (lines 09 and 13).

In the third sequence (lines 11-13) system misunderstands again the day suggested by the user (line 11), this time to indicate the timeslot as unavailable. Then the system asks again for a new timeslot (line 12) thus initiating the fourth and last sequence of the excerpt (lines 15-17). At this point the user starts performing less formatted actions like emotional expressions (boredom, irritation, laughter) that increase as the dialogue unfolds. For example, this system's turn is overlapped by the user expressing her boredom (breathing out loudly, line 13); then her misunderstood utterance produces an off talk at low voice (same line). After that, she laughs while answering the timeslot question once again (line 15). Nevertheless she follows the previous strategy that consists in choosing a different day rather than keeping on trying to get the timeslot by rewording the day description.

Time	Line	Spk	Transcription	Comments
	01	S	désolée (.) je ne suis pas sûre d'avoir bien compris (.) quand souhaitez-vous prendre rendez-vous ?	
0:55	02	U	mercredi quatorze juillet ? (.) au matin	
	03	S	merci de confirmer la réservation du créneau suivant, le mercredi vingt et un juillet avec le passage du technicien dans la matinée. est-ce que cela vous convient ?	S hears only the day, not the date
	04		(2.0)	
	05	U	NON	
	06	S	vous venez de refuser un créneau. aucun autre créneau ne satisfait vos contraintes. nous allons reprendre depuis le début. essayez de spécifier d'autres contraintes. quand souhaitez-vous prendre rendez-vous ?	
1:27	07	U	mardi treize juillet	
1:30	08	S	merci de confirmer la réservation du créneau suivant, le mardi vingt juillet avec le passage du technicien dans l'après-midi. est-ce que cela vous convient ?	S hears only the day, not the date[2]
1:41	09	U	NON	
	10	S	vous venez de refuser un créneau. aucun autre créneau ne satisfait vos contraintes. nous allons reprendre depuis le début. essayez de spécifier d'autres contraintes. quand souhaitez-vous prendre rendez-vous ?	
	11	U	jeudi vingt deux juillet. matin	U asks next Thursday
	12	S	vous avez une préférence pour jeudi de cette semaine dans la matinée. je suis désolée nous n'avons trouvé au[cun créneau satisfaisant vos contraintes (.) nous allons reprendre [depuis le début(.)	S hears only the day not the date and refers to this Thursday
	13	U	[hhh [°pourquoi xxx°	
	14	S	essayez de spécifier d'autres contraintes. (.) quand souhaitez-vous prendre rendez-vous ?	
	15	U	samedi vingt quatre juillet ((laughs)) le matin	
	16	S	vous avez une préférence pour samedi de cette semaine je suis désolée nous n'avons trouvé aucun créneau satisfaisant vos contraintes (.) nous allons [reprendre depuis le début(.)essayez de spécifier d'autres contr[aintes.	
	17	U	[hhhh [xxx	
	18	S	désolée. (.) je ne suis pas sûre d'avoir bien compris. avez-vous dit demain ?	

Transcript 5.4.2: Switching the day. (1013+ EXPER : 34374 - PE)

[2]The repetition of this partial understanding is explained by the fact that the user spoke every word in a separated way, probably thinking it would help the system's understanding.

Switching the day, seen as users' strategy to cope with the system's misunderstanding of a wished timeslot can be compared to the action to restart the computer as a troubleshooting solution. Actually, the user displays that she does not understand the machine's behaviour but accepts the interactional incoherences without trying to clarify and repair them. Building on acceptance of normality of system's irrelevant turns, she seeks to move on by trying something else. To that purpose, she is ready to continuously drop a given timeslot and select another one to negotiate.

This type of interactions manifest user's orientation to the presumption that in some cases the machine simply does not understand and produces irrelevant responses ("the machine is stupid"). But there's no way to grasp why this is so ("the machine's functioning is impenetrable") and to adapt her behaviour so there are no apparent reasons to engage in repair sequences.

To put it differently, users distrust system's competence to understand correctly her utterances and produce relevant answers. From users' point of view system's behaviour is framed by distrust, i.e. the system appears to be under the presumption of speech recognition error and interactional incompetence in general (on this point, see subsection 5.2.1).

One way to manage distrust is to move on by switching the day as we have seen. But the distrust attitude to the machine may give rise to very different users' actions: the persisting rewordings as error recovery strategy as we will see in the next paragraph.

Managing interactional misalignment: rewording the same day reference Transcript 5.4.3 below displays the same phenomenon of interactional misalignment due to the system ignoring complex time information. In the timeslot negotiation between the user and system the user refers to Thursday of the following week twice (lines 02 and 04), while the system is oriented to Thursday of the current week (lines 03 and 05) and does not identify the problem. In the face of this blind alley type of interaction some users will repeat exactly the same utterance; others will drop the timeslot and change the day (see previous paragraph). The excerpt below displays a very different form of users' action: rewording the same day (see lines 02, 04 and 06 for three different references to the same day). The second rewording (line 06) is correctly "heard" by the system (in line 08) and allows the interaction to move on from the blind alley model as the system does not repeat the previous message (as in lines 03 and 05) but asks for confirmation (line 08). On this basis the interaction can progress to the next stage, i.e. negotiating of a new timeslot (lines 10-11). That demonstrates that users' rewording practices are relevant error recovery actions that might lead to interactional realignment with the system. It is worth noting that the basic distrust attitude to the machine and the same basic presumption of potential speech recognition error and incompetence of the system give rise to very different users' actions: repetitions, rewordings and day switching practices.

Time	Line	Spk	Transcription	Comments
6:28	01	S	[essayez de spécifier d'autres contraintes (.) quand souhaitez-vous prendre rendez-vous/]	
6:41	02	U	[hhhhhhh] jeudi vingt deux hhh	Timeslot suggestion. U means Thursday of the next week

Time	Line	Spk	Transcription	Comments
	03	S	vous avez une préférence pou :r/ jeudi de cette semaine (.) je suis désolée nous n'avons trouvé aucun créneau satisfaisant vos contraintes (.) nous allons reprendre depuis le [début (.) essayez de spécifier d'autres contraintes (.) quand souhaitez-vous prendre rendez-vous/]	S suggests Thursday of this week
7:01	04	U	[hhhhhhh] jeudi prochain	1st rewording of line 02
	05	S	vous avez une préférence pou :r ? jeudi de cette semaine (.) je suis désolée nous n'avons trouvé aucun créneau satisfaisant vos contraintes (.) nous allons reprendre depuis le début (.) essayez de spécifier d'autres contraintes (.) quand souhaitez-vous prendre rendez-vous ?	S suggests once again Thursday of this week
7:18	06	U	le vingt deux hhhh.	2nd rewording of line 02
	07		(3.0)	
	08	S	désolée je ne suis pas sûre d'avoir bien compris (.) avez-vous dit le vingt deux juillet ? merci de confirmer par oui ou par non.	S asks for confirmation
7:30	09	U	¿oui¡/ hhh	U displays exhaustion
	10	S	vous avez une préférence pou :r ? le vingt deux juillet (.) je suis désolée nous n'avons trouvé aucun créneau satisfaisant vos contraintes (.) nous allons reprendre depuis le début (.) essayez de spécifier d'autres contraintes (.) quand souhaitez-vous prendre rendez-vous ?	
7:48	11	U	Hhh ? samedi vingt quatre hhh.	

Transcript 5.4.3: Rewording (1013+ EXPER : 89284 - PE)

Recommendation: Extend the ASR to words preceding [NEXT-THIS_DAY] and following [DATE] the day formulation defining most often the week in which this particular day is to be found. Encourage users to reword timeslot suggestions after the first repetition of the same day.

5.5 Concluding remarks: overall evaluation and predictive limits of the analysis

The overall effectiveness of the 1013+ is high: task completion is the rule and error recovery often takes place smoothly. Part of the global success of the system is due to the severe constrains it imposes on the user's expressiveness, which maximize the chance that recognition will succeed and that the dialogue manager will provide an appropriate response. In part this is achieved through questions that encourage answers from a comparatively limited set of words and expressions, e.g. the vocabulary of date and time references. But the most effective device in this regard is no doubt the binary yes/no question, which not only restrains the space of possible answers to two possibilities, but also minimizes the risk of recogni-

tion error by requiring that the user selects either of two markedly contrasted monosyllables. The binary question's success is proportional to its ability to make the user's answer predictable and easy to identify. In what follows, we evaluate the predictive power of the 1013+ experiment by focusing solely on some tricky contrasts regarding the way the yes/no binary question works in naturalistic and experimental settings. We do not address the implications of other apparent contrasts like task complexity, ambient noise, user competence, and the relationship between user satisfaction and task completion.

However, the binary question's effectiveness is not equally high in the naturalistic and experimental corpora. If one is tempted to ask why the success rates should differ, the answer appears to reveal striking differences between the way users set an appointment in a real-life situation and in the context of an experimental task. These differences are relevant to the extent that they set limits to the representativeness of the experimental results as a reliable projection of future real uses. Why, then, is the binary yes/no question less successful in real situations than in the experiment?

When one looks closely to the instructions and the structure of the task, one realises that the experiment relies on two fictions, the consequences of which are tremendously important for understanding the experiment's limited ability to simulate real-life uses. On the one hand, the user's agenda is made to the image of the system's: clearly delimited time slots distributed into two mutually exclusive classes, namely available and not available periods of the day and of the week. The remark that this is quite an unusual way of organising one's agenda might look trivial at first sight, but it is not. Two observations are in order: first, the distribution into available and not available slots is flexible and therefore revisable in real life. You are usually at work on Monday morning; this means that you are not available for things like going for a walk, which you use to do on Sunday morning. But if you are informed that a close relative has suffered an accident, you might not hesitate in being available on Monday morning, even if you are usually expected to be at work at that time. This is a dramatic example, but the same can be said of a trouble with your telephone line. If recovering the line is something urgent for you, which is the case of a number of real calls we have examined, you might be ready to change your agenda so that the technician solve the problem as soon as possible. Imagine you are supposed to be at work on Monday morning, but that turns out to be the only available slot of the telephone company's technician; if repairing the land line is really urgent for you, you might consider arriving later to work. Let us call this assumptionof the experimental setting *the inflexibility assumption*.

Apart from an inflexible agenda, the experimental device relies on another fiction: that the person who calls the system and sets the appointment is the person that will deal with the technician's visit. It would be weird to expect the dialogue system to come home to repair the broken line; similarly, one should not take for granted that the caller will be the one who will receive the technician. As a matter of fact, as we have seen the naturalistic corpus contains a number of dialogues in which the caller needs to check with somebody else at home before suggesting or accepting an appointment. The line is a single one, but the household hosts several people, among which a division of labour might be in place between the one who calls the system and the one who receives the technician. The experiment's assumptionconsists in stipulating that the caller is the receiver; let us call this assumption*the uniqueness assumption*.

When one considers the inflexibility and the uniqueness assumptions that are involved in the experiment, an explanation to the lower success rate of the yes/no question in the naturalistic corpus becomes available. The binary yes/no question is perfectly suited to an agenda that is structured exactly in the same way, i.e. in binary terms. As we saw, in the experiment the binary character of the system's yes/no question finds its logical reflection in the binary character of the user's distribution of time slots into available/not available. In view of her experimental agenda, the user can always tell whether she is or not available for a given time slot. In other words, every user can meaningfully say "yes" or "no", because the only two relevant

expanded messages are “yes, that time slot is available in my agenda ” or “no, that time slot is not available in my agenda ”. This simplification, which increases the system’s effectiveness, might not faithfully reflect real-life conditions. For people do not simply have available or unavailable time slots in their agendas; what they fundamentally have are projects, things to do in life, priorities, and problems to solve. The agenda is an approximate reminder of how available time should be distributed into these various projects, but the relationship between the agenda and what people really end up doing is always flexible. This fact, which seems so trivial, might be a cause of trouble in user-system real dialogues and is certainly the reason why the yes/no question does not always receive a yes/no answer. An appointment is not only the kind of thing one is available or unavailable for, it is also the kind of thing one might expect to take place at a certain point in time. Not only availability, but position in time is also relevant in fixing an appointment. If you have a terrible toothache, you do not want the dentist to tell you when she is available in the abstract; you want the dentist to tell you when she is available earliest. The naturalistic corpus shows that users do reason in this way, which eventually gets expressed in answers to the yes/no question that are surprisingly not couched in yes/no terms, e.g. “it’s too late” or “it’s urgent”. Conversely in the experimental setting, the inflexibility assumption, which reduces the agenda to a rigid binary scheme, rules out this possibility in advance. That contributes to higher recognition and task completion rates, but limits the experiment’s power to predict real-life uses. In short, users appear to be concerned not only with availability but most importantly with urgency. Precisely because what matters the most is time, real users might change their agendas in order to have the problem solved as soon as possible. The inflexibility assumption overlooks these commonplace facts of everyday life.

The uniqueness assumption, in turn, fuses the person that calls the system with the person that will receive the technician. The assumptionworks in the experimental setting, because the user’s inflexible agenda turns out to be at the same time the caller’s inflexible agenda and the receiver’s inflexible agenda. However, as the naturalistic corpus suggests, caller and receiver might be two different persons, with their respective flexible agendas. Again, this trivial fact might cause trouble in the dialogue, since the caller might need to check with the receiver if the latter is available at the suggested date and time. A number of annoying time-out messages in the naturalistic corpus is due to this practice. The possibility of these time-outs taking place in the experiment are ruled out in advance, simply because, being one and the same person, the caller will never need to consult the receiver. Again, this simplification increases recognition and success rates, but makes abstraction of a recurrent problem that the real commercial system will have to face.

To sum up, the overall evaluation of the 1013+ system is substantially positive, but one should be reminded that the inflexibility and uniqueness assumptions set certain limits to the predictive power of the experimental results.

Chapter 6

Conclusion

This document reported on the final evaluations of the CLASSiC TownInfo and Appointment Scheduling systems. We described the setup and results of the experiments involving real users calling different systems to perform different tasks and give ratings to each dialogue.

Part I of the report concerned the TownInfo system (System 1) and Part II concerned the Appointment Scheduling systems (Systems 2, 3, and 4) This report also presented the qualitative evaluation of the Appointment Scheduling systems carried out by France Telecom / Orange Labs (Part II, Chapter 5).

For the TownInfo systems, the actual domain was switched from an imaginary town to real locations in Cambridge and VoIP technology was used during evaluation. Subjects were asked to find a place to eat in Cambridge, following a scenario given to them. For the TownInfo evaluations a total of 2046 dialogues were collected and analysed.

For the Appointment Scheduling systems, the subjects were asked to book an appointment on one of the free slots in a user calendar given to them. Systems built by France Telecom and the academic partners were evaluated on the same tasks.

For both TownInfo and Appointment Scheduling (AS) domains, one of the evaluated systems included components contributed by different sites within the consortium. For more details about these integrated systems, see deliverable D5.2.2 for the CLASSiC TownInfo system, and deliverable D5.4 for the CLASSiC Appointment Scheduling system.

For the AS systems, System 2 collected a total of 628 dialogues, while Systems 3 and 4 collected 740 and 709 dialogues for evaluation respectively, for a total of 2077 AS dialogues.

6.1 The TownInfo System (System 1)

The main contrasts explored in these evaluations were the effects of processing N-best lists as input to the dialogue system (using POMDP techniques) as opposed to using only 1-best ASR input, and the effects of using the trained NLG component.

Even with average WERs over 50%, as shown in table 2.3.4, subjective success rates of 60% to 65% were achieved in the Feb'11 trial. This shows that the system was fairly resilient even when operating in extremely hostile conditions. Also, the improved HIS state space representation and pruning algorithms worked well enabling the systems to support prolonged dialogues without noticeable degradation in real time performance.

Whilst the evaluation results for both the Nov'10 and Feb'11 trials demonstrate the robustness of the systems in severe conditions, the overall performance was rather poor, chiefly we believe due to the poor ASR performance. As noted in section 2.3.1, when the data collected in the CamInfo trial was used to retrain the recogniser, the word error rate approximately halved and the dialogue success rate increased by over 20%.

In the Nov '10 and Feb'11 MTurk evaluations, the partial completion score for the CLASSiC system (i.e. including the trained NLG component) was significantly higher than the TownInfo system (p=0.02, z-test), suggesting that the more elaborate venue offers from the trained NLG component helped the user find the venue they were looking for more easily. However, from figure 2.3.1 and figure 2.3.3, we see that the CLASSiC TownInfo system (i.e. containing the trained NLG component) appeared to be more fragile than the other systems at high word error rates, especially in the MTurk subjective evaluation. This may be a consequence of trying to provide too much information to the user based on incorrect assumptions, which suggests that if the system is unsure, it should focus on offering a single entity but when confidence is high, the more intelligent presentation of information generated by the CLASSIC NLG system works well.

We also note that in one case, the Cambridge based Feb'11 evaluation, the inferred goal based success rates for the N-Best systems are better than those of the 1-Best system, and the success rate for the N-Best-CLASSiC system is significantly higher than for the 1-Best-UCAM system (p=0.03).

From table 2.3.4 we observe no other statistically significant improvements of the N-best system over the 1-best system. This may be due to poor a match between the simulator's error model and real data, which would impact on the N-best policy more than the 1-best policy.

However, we should emphasise that the similar performance of 1-best and N-best systems in this case does not mean that the POMDP framework is ineffective, or that an MDP would have worked just as well. In fact, POMDP systems are fundamentally different from MDP systems because POMDPs integrate information over both time and the N-best alternatives, whereas an MDP simply tracks the most likely state.

For a more detailed discussion of these results, please see section 2.4.

6.2 The Appointment Scheduling Systems

For the AS systems, objective task completion of all systems is high. Small differences between systems were visible. Those small differences were not necessarily linked to a chosen approach, but may be attributed to side-effects of local design differences. We would like to recall in this conclusion that System 3[1] was already the result of an on-line optimisation, which triggered a 10% task completion increase. This means that Systems' 2 and 4 performances already exceed classical handcrafted performance and thus that the project delivered three systems that are beyond the state of the art.

We note that System 3 also used non-commercial ASR models (the same as System 4's). Nevertheless, despite a strong observed WER, all systems achieved similar high-level objective task completion rates, of around 80%. This shows that the systems were quite robust even when operating in hostile conditions. All of these systems were developed rapidly and efficiently using the methods and tools developed during the CLASSiC project (see Section 3.2).

In the AS System 2, we also showed that the trained NLG component for Temporal Referring Expres-

[1] System 3 was used as a baseline.

sions brings significant benefits in terms of users' perceived task completion (+23.7%), and overall user satisfaction (+5%), together with shorter Call Duration in terms of time (-15.7%) and average number of words per system turn (-23.93%) [20].

A further finding of this experimentation is that some users seem to appreciate being directed, influenced, or constrained during their dialogue. These users are looking for an efficient, frustration-free spoken dialogue system that minimises interpretation (ASR and SLU) errors and rejects. It is counter-productive (for such users) to allow too much interaction in order to produce more 'natural' dialogues. Indeed, the subjective overall rating of heavily-directed System 3 is higher than that of Systems 2 and 4, which employed more open questions such as "when are you available?".

A sociological study presented further detailed qualitative analysis of the AS dialogues using methods from Conversation Analysis, for example examining types of errors and interactional misalignment phenomena between the user and the system. This study was based on the CLASSiC final experimentation but also on the commercial application corpus. This study leads to interesting observations, such as the slight differences in user behaviour between the experimental and commercial corpora, and to suggestions for improvements of the systems, such as adapting better to the temporal references used by the users.

Finally, taken together, the experimentation results demonstrate that the statistical learning methods and tools developed in the CLASSiC project provide a promising foundation for future research and development into robust and adaptive spoken dialogue systems.

Bibliography

[1] S. Young, M. Gašić, S. Keizer, F. Mairesse, B. Thomson, and K. Yu. The Hidden Information State model: a practical framework for POMDP based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2009.

[2] B. Thomson, K. Yu, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, and S. Young. Evaluating semantic-level confidence scores with multiple hypotheses. In *Proceedings Interspeech*, Brisbane, Australia, September 2008.

[3] G. Putois, R. Laroche, and P. Bretier. Online reinforcement learning for spoken dialogue systems: The story of a commercial deployment success. In *Proceedings of SIGDIAL*, Tokyo (Japan), September 2010.

[4] P. Bretier, R. Laroche, and G. Putois. D5.3.4: Industrial self-help system ("system 3") adapted to final architecture. Prototype D5.3.4, CLASSIC Project, 2010.

[5] R. Laroche and G. Putois. D5.5: Advanced appointment-scheduling system "system 4". PRototype D5.5, CLASSIC Project, 2010.

[6] Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*, 6, 2000.

[7] R. Laroche, B. Bouchon-Meunier, and P. Bretier. Uncertainty management in dialogue systems. In *Proceedings of the European Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2008.

[8] Sebastian Möller and Nigel G. Ward. A framework for model-based evaluation of spoken dialog systems. In *Proc. of SIGDIAL*, 2008.

[9] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proc. of ACL*, 2001.

[10] Gabriel Skantze and Anna Hjalmarsson. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 1–8, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] Silvan Heintze, Timo Baumann, and David Schlangen. Comparing local and sequential models for statistical incremental natural language understanding. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 9–16, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[12] Harold Garfinkel. *Studies in ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ [u.a.], 1. print. edition, 1967.

[13] Harvey Sacks. *Lectures on Conversation*. Wiley-Blackwell; Volumes I and II edition, 1974.

[14] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735, 1974.

[15] Emanuel A. Schegloff. *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. Cambridge University Press, January 2007.

[16] J. Maxwell Atkinson. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, 1985.

[17] George Psathas. *Conversation Analysis: The Study of Talk-in-Interaction*. Sage Publications, Inc, 1995.

[18] N. Frijda. The laws of emotion. *Cognition and Emotion*, 1:235–258, 1988.

[19] K. Scherer. *Appraisal Processes in Emotion: Theory, Methods, Research (Series in Affective Science)*. Oxford University Press, USA, February 2001.

[20] Srinivasan Janarthanam, Helen Hastie, Oliver Lemon, and Xingkun Liu. 'The day after the day after tomorrow? ' A machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *Proceedings of SIGDIAL*, 2011. (to appear).

Appendix A

Appointment scheduling statistics

Systems 2&3&4	Sys. task completion				Q. task completion				Call duration				Number of ASR rejects				Q. ASR rating				Q. phrasing rating				Q. TTS rating				Q. overall rating				Q. task ease				Q. future use				Q. number of call				Volume			
Mean	0,81	0,84	0,80	0,79	0,79	0,82	0,78	0,77	87	77	90	94	0,54	0,56	0,48	0,58	4,58	4,76	4,59	4,40	5,45	5,50	5,43	5,41	4,79	4,87	4,76	4,76	6,45	6,82	6,42	6,10	4,09	4,39	4,05	3,83	3,90	4,10	3,88	3,71	2,12	2,23	2,05	2,07	2077	707	651	719
	0,85	0,83	0,86	0,88	0,82	0,81	0,81	0,85	78	79	75	79	0,49	0,65	0,31	0,36	4,80	4,77	4,87	4,78	5,49	5,50	5,49	5,48	4,83	4,85	4,80	4,83	6,88	6,92	6,86	6,83	4,39	4,43	4,35	4,34	4,09	4,08	4,08	4,12	2,17	2,31	1,93	2,16	705	347	180	178
	0,81	0,82	0,82	0,80	0,79	0,80	0,78	0,79	89	82	76	97	0,53	0,59	0,14	0,61	4,52	4,59	4,62	4,45	5,44	5,47	5,41	5,44	4,82	4,93	4,82	4,77	6,32	6,50	6,45	6,20	4,00	4,11	4,32	3,86	3,83	3,87	4,02	3,76	2,08	2,18	2,23	1,99	625	177	102	346
	0,78	0,87	0,78	0,69	0,76	0,85	0,77	0,67	94	67	101	104	0,60	0,36	0,65	0,72	4,43	4,91	4,44	3,98	5,41	5,54	5,41	5,30	4,74	4,85	4,72	4,68	6,14	6,96	6,20	5,26	3,88	4,60	3,82	3,32	3,77	4,35	3,75	3,26	2,10	2,14	2,05	2,13	747	183	369	195
95% window	0,02	0,03	0,03	0,03	0,02	0,03	0,03	0,03	3	5	4	4	0,06	0,11	0,09	0,09	0,06	0,11	0,11	0,11	0,04	0,06	0,07	0,07	0,05	0,09	0,09	0,08	0,12	0,20	0,21	0,20	0,09	0,15	0,16	0,14	0,08	0,14	0,15	0,14	0,07	0,12	0,13	0,12				
	0,03	0,04	0,05	0,05	0,03	0,04	0,06	0,05	5	8	8	8	0,11	0,20	0,11	0,12	0,11	0,16	0,21	0,21	0,06	0,09	0,13	0,12	0,09	0,13	0,18	0,16	0,20	0,29	0,38	0,38	0,15	0,22	0,30	0,28	0,15	0,21	0,29	0,28	0,12	0,18	0,24	0,24				
	0,03	0,06	0,08	0,04	0,03	0,06	0,08	0,04	4	9	7	6	0,09	0,18	0,11	0,12	0,12	0,23	0,29	0,15	0,07	0,13	0,19	0,09	0,09	0,16	0,21	0,12	0,21	0,40	0,51	0,27	0,15	0,31	0,38	0,19	0,15	0,30	0,38	0,19	0,13	0,25	0,34	0,17				
	0,03	0,05	0,04	0,07	0,03	0,05	0,04	0,07	4	7	6	8	0,10	0,13	0,13	0,23	0,11	0,19	0,15	0,23	0,07	0,12	0,09	0,15	0,08	0,17	0,11	0,17	0,20	0,36	0,28	0,41	0,14	0,26	0,20	0,29	0,14	0,25	0,20	0,29	0,12	0,24	0,16	0,22				

System 2	Sys. task completion				Q. task completion				Call duration				Number of ASR rejects				Q. ASR rating				Q. phrasing rating				Q. TTS rating				Q. overall rating				Q. task ease				Q. future use				Q. number of call				Volume			
Mean	0,79	0,84	0,78	0,76	0,68	0,74	0,67	0,64	97	82	98	110	0,00	0,00	0,00	0,00	4,15	4,34	4,12	4,02	5,10	5,09	5,10	5,11	4,23	4,21	4,26	4,22	5,21	5,70	5,23	4,75	3,67	4,14	3,57	3,35	3,44	3,66	3,41	3,25	2,32	2,60	2,24	2,14	628	197	221	210
	0,84	0,82	0,83	0,88	0,71	0,72	0,67	0,76	85	82	78	101	0,00	0,00	0,00	0,00	4,27	4,24	4,32	4,24	5,16	5,14	5,18	5,16	4,21	4,22	4,20	4,20	5,54	5,65	5,53	5,31	3,91	4,19	3,75	3,53	3,51	3,64	3,47	3,31	2,35	2,70	2,00	2,08	207	98	60	49
	0,81	0,83	0,75	0,83	0,71	0,76	0,68	0,71	101	91	89	113	0,00	0,00	0,00	0,00	4,18	4,30	4,06	4,20	5,09	4,96	5,09	5,16	4,35	4,28	4,51	4,29	5,27	5,43	5,47	5,07	3,65	3,87	3,62	3,56	3,50	3,59	3,36	3,53	2,39	2,61	2,58	2,19	195	46	53	96
	0,73	0,89	0,76	0,57	0,62	0,75	0,66	0,46	103	73	113	112	0,00	0,00	0,00	0,00	4,03	4,53	4,05	3,58	5,06	5,09	5,06	5,02	4,15	4,11	4,18	4,14	4,87	6,00	4,94	3,85	3,48	4,26	3,44	2,89	3,31	3,77	3,41	2,78	2,22	2,40	2,19	2,12	226	53	108	65
95% window	0,03	0,05	0,06	0,06	0,04	0,06	0,06	0,07	5	9	8	8	0,00	0,00	0,00	0,00	0,11	0,20	0,19	0,19	0,09	0,15	0,15	0,15	0,09	0,16	0,15	0,15	0,23	0,41	0,37	0,39	0,16	0,28	0,27	0,27	0,15	0,26	0,26	0,26	0,14	0,25	0,23	0,23				
	0,05	0,08	0,10	0,09	0,06	0,09	0,12	0,12	9	14	15	14	0,00	0,00	0,00	0,00	0,20	0,29	0,40	0,37	0,15	0,20	0,29	0,30	0,16	0,25	0,32	0,29	0,40	0,61	0,74	0,74	0,28	0,40	0,57	0,55	0,27	0,37	0,55	0,50	0,25	0,38	0,45	0,48				
	0,06	0,11	0,12	0,08	0,06	0,13	0,13	0,09	9	19	12	14	0,00	0,00	0,00	0,00	0,20	0,44	0,40	0,26	0,15	0,33	0,32	0,20	0,16	0,28	0,32	0,23	0,40	0,84	0,77	0,57	0,28	0,59	0,59	0,37	0,27	0,53	0,55	0,37	0,24	0,49	0,46	0,35				
	0,06	0,09	0,08	0,12	0,06	0,12	0,09	0,12	8	13	11	15	0,00	0,00	0,00	0,00	0,18	0,36	0,25	0,35	0,15	0,31	0,21	0,28	0,14	0,30	0,20	0,27	0,37	0,68	0,52	0,70	0,26	0,53	0,35	0,51	0,25	0,49	0,34	0,49	0,22	0,46	0,31	0,43				

System 3	Sys. task completion				Q. task completion				Call duration				Number of ASR rejects				Q. ASR rating				Q. phrasing rating				Q. TTS rating				Q. overall rating				Q. task ease				Q. future use				Q. number of call				Volume			
Mean	0,81	0,82	0,84	0,77	0,83	0,82	0,86	0,81	69	54	70	83	0,26	0,17	0,21	0,39	5,10	5,37	5,20	4,73	5,69	5,79	5,70	5,58	5,11	5,15	5,10	5,07	7,40	7,76	7,58	6,85	4,62	4,92	4,80	4,14	4,32	4,49	4,56	3,91	2,00	2,06	2,00	1,94	740	256	237	247
	0,83	0,83	0,80	0,87	0,85	0,83	0,82	0,90	63	52	74	73	0,15	0,10	0,16	0,21	5,34	5,44	5,34	5,16	5,69	5,72	5,69	5,64	5,12	5,04	5,20	5,19	7,84	7,95	7,72	7,76	4,83	4,95	4,69	4,76	4,49	4,50	4,44	4,51	2,10	2,19	2,00	2,01	248	120	61	67
	0,78	0,79	0,89	0,72	0,80	0,77	0,89	0,78	73	62	63	86	0,29	0,29	0,28	0,30	4,91	5,10	5,20	4,66	5,73	5,89	5,78	5,61	5,13	5,24	5,15	5,05	7,07	7,33	7,48	6,71	4,36	4,57	5,04	3,92	4,07	4,16	4,70	3,73	1,85	1,97	1,78	1,80	220	70	46	104
	0,81	0,86	0,83	0,74	0,84	0,86	0,86	0,78	71	51	70	89	0,33	0,18	0,21	0,68	5,04	5,53	5,13	4,45	5,65	5,80	5,68	5,49	5,07	5,24	5,04	4,99	7,26	7,88	7,55	6,22	4,63	5,24	4,76	3,88	4,37	4,83	4,57	3,63	2,03	1,91	2,08	2,05	272	66	130	76
95% window	0,03	0,05	0,05	0,05	0,03	0,05	0,05	0,05	3	6	4	5	0,06	0,08	0,08	0,14	0,10	0,16	0,16	0,20	0,05	0,07	0,08	0,10	0,07	0,13	0,12	0,12	0,17	0,29	0,26	0,30	0,14	0,26	0,21	0,23	0,14	0,25	0,22	0,24	0,11	0,19	0,20	0,19				
	0,05	0,07	0,10	0,08	0,05	0,07	0,10	0,07	6	8	12	9	0,06	0,06	0,13	0,12	0,15	0,21	0,32	0,31	0,09	0,12	0,18	0,17	0,13	0,19	0,25	0,22	0,27	0,41	0,51	0,49	0,25	0,38	0,48	0,40	0,25	0,38	0,46	0,43	0,20	0,29	0,41	0,36				
	0,06	0,10	0,09	0,09	0,05	0,10	0,09	0,08	6	15	4	6	0,11	0,19	0,23	0,15	0,20	0,36	0,35	0,31	0,08	0,09	0,14	0,15	0,13	0,24	0,25	0,18	0,31	0,62	0,54	0,43	0,26	0,54	0,40	0,35	0,26	0,54	0,43	0,35	0,22	0,38	0,50	0,30				
	0,05	0,08	0,07	0,10	0,04	0,08	0,06	0,10	5	11	4	10	0,13	0,19	0,11	0,38	0,17	0,25	0,22	0,37	0,08	0,11	0,10	0,20	0,12	0,26	0,17	0,22	0,29	0,52	0,37	0,60	0,21	0,40	0,27	0,44	0,22	0,39	0,30	0,46	0,18	0,36	0,26	0,35				

System 4	Sys. task completion				Q. task completion				Call duration				Number of ASR rejects				Q. ASR rating				Q. phrasing rating				Q. TTS rating				Q. overall rating				Q. task ease				Q. future use				Q. number of call				Volume			
Mean	0,83	0,85	0,80	0,83	0,85	0,88	0,82	0,84	98	96	107	93	1,31	1,38	1,35	1,22	4,42	4,48	4,37	4,40	5,51	5,54	5,49	5,50	4,97	5,09	4,90	4,90	6,54	6,76	6,36	6,47	3,91	4,06	3,67	3,93	3,87	4,04	3,60	3,90	2,07	2,12	1,89	2,15	709	254	193	262
	0,88	0,84	0,93	0,89	0,88	0,86	0,93	0,89	88	103	75	69	1,23	1,65	0,76	0,81	4,70	4,55	4,95	4,77	5,58	5,57	5,59	5,56	5,06	5,14	5,00	4,94	7,03	6,91	7,31	7,02	4,34	4,13	4,61	4,53	4,18	4,03	4,32	4,35	2,10	2,11	1,80	2,37	250	129	59	62
	0,84	0,85	1,00	0,84	0,86	0,87	1,00	0,86	95	98	63	94	1,26	1,38	0,33	1,23	4,40	4,21	5,67	4,46	5,47	5,38	5,33	5,51	4,94	5,05	5,33	4,88	6,53	6,36	8,00	6,57	3,95	3,75	5,67	4,00	3,90	3,75	5,33	3,92	2,04	2,10	2,67	2,01	210	61	3	146
	0,77	0,86	0,73	0,76	0,81	0,92	0,77	0,76	110	80	123	117	1,43	0,84	1,63	1,65	4,14	4,58	4,08	3,80	5,48	5,64	5,44	5,37	4,90	5,05	4,84	4,89	6,06	6,81	5,89	5,59	3,43	4,20	3,21	3,04	3,53	4,33	3,23	3,30	2,06	2,17	1,92	2,26	249	64	131	54
95% window	0,03	0,04	0,06	0,05	0,03	0,04	0,05	0,04	5	9	10	7	0,13	0,27	0,23	0,19	0,11	0,18	0,21	0,19	0,06	0,09	0,11	0,10	0,08	0,13	0,16	0,14	0,18	0,30	0,35	0,30	0,14	0,23	0,30	0,22	0,14	0,23	0,29	0,22	0,12	0,20	0,23	0,19				
	0,04	0,06	0,07	0,08	0,04	0,06	0,07	0,08	10	15	16	15	0,27	0,47	0,28	0,27	0,19	0,26	0,33	0,40	0,08	0,12	0,16	0,17	0,14	0,19	0,29	0,29	0,30	0,43	0,54	0,61	0,24	0,35	0,46	0,47	0,23	0,33	0,45	0,45	0,20	0,28	0,40	0,41				
	0,05	0,09	0,00	0,06	0,05	0,09	0,00	0,06	6	11	2	8	0,20	0,39	0,54	0,23	0,19	0,35	0,54	0,23	0,11	0,22	1,09	0,13	0,14	0,25	1,09	0,17	0,32	0,58	1,89	0,39	0,22	0,43	0,54	0,26	0,23	0,45	1,09	0,27	0,22	0,41	1,44	0,26				
	0,05	0,09	0,08	0,12	0,05	0,07	0,07	0,12	8	10	12	16	0,21	0,26	0,29	0,56	0,19	0,34	0,26	0,44	0,10	0,15	0,14	0,29	0,14	0,27	0,19	0,35	0,31	0,58	0,43	0,69	0,25	0,40	0,36	0,50	0,25	0,40	0,36	0,53	0,20	0,41	0,29	0,39				

Figure 1: Systems evaluation: mean and 95% window for the key performance indicators (see section 3.3 for details).

Systems 2&3&4	Sys. task completion	Q. task completion	Call duration	Number of ASR rejects	Q. ASR rating	Q. phrasing rating	Q. TTS rating	Q. overall rating	Q. task ease	Q. future use
Q. task completion	0,76 0,79 0,77									
	0,75 0,77 0,72 0,76									
	0,79 0,84 0,82 0,76									
	0,78 0,66 0,82 0,77									
Call duration	-0,23 -0,37 -0,24 -0,07 -0,33 -0,44 -0,31 -0,22									
	-0,31 -0,36 -0,33 -0,14 -0,39 -0,45 -0,36 -0,27									
	-0,14 -0,43 -0,03 0,01 -0,24 -0,47 -0,05 -0,15									
	-0,21 -0,29 -0,22 -0,06 -0,34 -0,39 -0,34 -0,23									
Number of ASR rejects	-0,17 -0,16 -0,21 -0,16 -0,14 -0,15 -0,15 -0,12	0,45	0,51 0,49 0,34							
	-0,13 -0,14 -0,08 -0,18 -0,12 -0,14 -0,02 -0,14	0,49	0,54 0,41 0,32							
	-0,14 -0,20 0,12 -0,15 -0,11 -0,20 0,13 -0,10	0,33	0,43 0,06 0,28							
	-0,22 -0,22 -0,26 -0,14 -0,17 -0,11 -0,22 -0,10	0,48	0,44 0,53 0,42							
Q. ASR rating	0,35 0,36 0,30 0,37 0,44 0,46 0,41 0,45	-0,39	-0,46 -0,35 -0,33	-0,29	-0,34 -0,26 -0,27					
	0,35 0,39 0,24 0,41 0,47 0,49 0,46 0,44	-0,42	-0,45 -0,32 -0,50	-0,31	-0,35 -0,25 -0,32					
	0,35 0,47 0,41 0,27 0,44 0,52 0,49 0,38	-0,32	-0,52 -0,17 -0,23	-0,28	-0,38 -0,13 -0,25					
	0,33 0,16 0,29 0,42 0,42 0,31 0,37 0,51	-0,39	-0,45 -0,37 -0,28	-0,28	-0,29 -0,28 -0,24					
Q. phrasing rating	0,08 0,06 0,14 0,06 0,18 0,15 0,23 0,16	-0,15	-0,16 -0,13 -0,16	-0,02	-0,01 -0,03 -0,04	0,34	0,35 0,40 0,29			
	0,12 0,10 0,19 0,10 0,25 0,23 0,39 0,13	-0,14	-0,14 -0,08 -0,23	-0,04	-0,01 -0,07 -0,19	0,39	0,33 0,47 0,43			
	0,01 0,00 0,14 -0,03 0,14 0,09 0,22 0,14	-0,16	-0,19 -0,04 -0,18	0,00	-0,03 0,08 0,00	0,30	0,36 0,44 0,23			
	0,11 0,02 0,11 0,12 0,16 0,03 0,16 0,18	-0,15	-0,16 -0,16 -0,08	-0,02	0,06 -0,03 0,00	0,33	0,36 0,35 0,25			
Q. TTS rating	0,11 0,12 0,09 0,13 0,18 0,20 0,16 0,16	-0,13	-0,09 -0,13 -0,16	0,02	0,07 -0,02 -0,03	0,36	0,36 0,33 0,37 0,39			
	0,09 0,13 -0,03 0,13 0,19 0,24 0,14 0,12	-0,08	-0,05 -0,03 -0,24	0,04	0,07 -0,01 -0,05	0,37	0,36 0,31 0,48 0,39			
	0,10 0,16 0,09 0,07 0,17 0,22 0,20 0,14	-0,13	-0,22 0,17 -0,14	0,00	0,00 0,04 -0,01	0,37	0,42 0,30 0,36 0,40			
	0,14 0,05 0,14 0,19 0,17 0,12 0,16 0,21	-0,17	-0,08 -0,23 -0,11	0,01	0,17 -0,01 -0,03	0,34	0,35 0,35 0,29 0,37			
Q. overall rating	0,45 0,44 0,44 0,45 0,57 0,56 0,56 0,58	-0,42	-0,46 -0,38 -0,37	-0,21	-0,24 -0,20 -0,18	0,74	0,73 0,73 0,74 0,40	0,40 0,43 0,36 0,45	0,44 0,42 0,48	
	0,41 0,42 0,36 0,46 0,56 0,55 0,61 0,55	-0,41	-0,44 -0,32 -0,45	-0,21	-0,25 -0,16 -0,20	0,78	0,77 0,77 0,80 0,43	0,40 0,53 0,41 0,49	0,45 0,49 0,57	
	0,41 0,55 0,47 0,33 0,54 0,63 0,55 0,50	-0,34	-0,49 -0,16 -0,28	-0,16	-0,22 -0,11 -0,13	0,70	0,70 0,70 0,69 0,36	0,35 0,47 0,33 0,43	0,42 0,29 0,47	
	0,49 0,35 0,46 0,56 0,59 0,51 0,53 0,68	-0,46	-0,53 -0,44 -0,39	-0,23	-0,27 -0,22 -0,20	0,72	0,69 0,70 0,73 0,39	0,44 0,38 0,35 0,43	0,46 0,42 0,43	
Q. task ease	0,49 0,55 0,52 0,41 0,66 0,72 0,67 0,61	-0,57	-0,60 -0,57 -0,51	-0,31	-0,34 -0,35 -0,26	0,67	0,68 0,66 0,64 0,30	0,30 0,32 0,28 0,29	0,26 0,29 0,30	0,70 0,69 0,71 0,70
	0,56 0,60 0,54 0,53 0,74 0,75 0,76 0,70	-0,58	-0,58 -0,54 -0,64	-0,29	-0,35 -0,21 -0,27	0,66	0,66 0,64 0,69 0,37	0,34 0,43 0,36 0,28	0,27 0,19 0,42	0,68 0,65 0,70 0,75
	0,44 0,60 0,49 0,34 0,62 0,74 0,69 0,54	-0,50	-0,64 -0,24 -0,46	-0,25	-0,34 -0,04 -0,22	0,65	0,72 0,76 0,57 0,30	0,26 0,36 0,31 0,28	0,27 0,30 0,28	0,68 0,72 0,69 0,65
	0,46 0,35 0,52 0,39 0,63 0,63 0,62 0,61	-0,60	-0,63 -0,63 -0,45	-0,38	-0,36 -0,43 -0,29	0,67	0,68 0,64 0,65 0,24	0,26 0,26 0,16 0,29	0,25 0,33 0,22	0,73 0,77 0,71 0,70
Q. future use	0,48 0,50 0,50 0,45 0,65 0,68 0,64 0,63	-0,46	-0,48 -0,47 -0,43	-0,25	-0,24 -0,29 -0,23	0,63	0,65 0,63 0,62 0,31	0,33 0,32 0,28 0,36	0,36 0,34 0,38	0,72 0,73 0,70 0,72 0,87
	0,52 0,55 0,47 0,53 0,71 0,71 0,70 0,71	-0,47	-0,45 -0,47 -0,53	-0,21	-0,24 -0,18 -0,18	0,66	0,68 0,62 0,65 0,38	0,37 0,42 0,34 0,38	0,39 0,29 0,47	0,71 0,69 0,71 0,75 0,88
	0,47 0,58 0,52 0,39 0,63 0,73 0,67 0,57	-0,39	-0,55 -0,20 -0,32	-0,20	-0,28 -0,03 -0,18	0,62	0,65 0,74 0,56 0,29	0,25 0,37 0,29 0,34	0,34 0,28 0,36	0,70 0,73 0,70 0,69 0,86
	0,46 0,25 0,50 0,44 0,61 0,55 0,60 0,63	-0,51	-0,47 -0,52 -0,44	-0,31	-0,19 -0,36 -0,27	0,61	0,57 0,59 0,62 0,26	0,34 0,26 0,19 0,36	0,35 0,38 0,32	0,73 0,80 0,69 0,70 0,87
Q. number of call	0,03 0,07 0,00 0,03 -0,01 0,02 -0,01 -0,03	-0,05	-0,04 -0,06 -0,02	-0,07	-0,06 -0,05 -0,10	0,02	0,02 -0,03 0,05 -0,05	0,00 -0,11 -0,06 0,01	-0,02 0,03 0,01	0,00 0,04 -0,04 -0,01 -0,01
	-0,02 -0,02 -0,08 0,08 -0,08 -0,07 -0,13 -0,04	0,01	0,01 0,02 0,00 -0,03	-0,03	-0,03 -0,02 -0,14	-0,02	-0,04 -0,08 0,09 -0,05	-0,01 -0,11 -0,06 -0,05	-0,04 -0,07 -0,06	-0,03 -0,03 -0,05 -0,03 -0,07
	0,05 0,16 -0,03 0,02 -0,01 0,12 -0,13 -0,05	-0,03	-0,11 -0,13 0,06	-0,10	-0,16 -0,09 -0,06	0,05	0,11 -0,02 0,04 -0,04	0,03 -0,13 -0,04	0,04 -0,01 0,05	0,13 -0,17 -0,04 -0,04
	0,06 0,18 0,04 0,02 0,06 0,09 0,08 0,01	-0,12	-0,15 -0,09 -0,16	-0,09	-0,07 -0,06 -0,15	0,02	0,07 0,00 0,02 -0,08	0,03 -0,11 -0,08	0,04 -0,04 0,09	0,02 0,10 0,01 0,05 0,07

Q. task ease	Q. future use
0,87	0,89 0,87
0,87	0,90 0,89
0,87	0,88 0,85
0,85	0,88 0,86
-0,02	-0,02 0,00
-0,10	-0,10 0,01
-0,07	-0,07 -0,07 0,02
-0,04	-0,04 -0,07 -0,10 -0,01
0,09	0,03 0,13 0,09
0,04	-0,04 -0,07 -0,10 -0,01 0,11 -0,06 -0,06
0,09	0,09 0,09 0,10

Figure 2: Systems evaluation: Correlations between key performance indicators (see section 3.3 for details) in aggregate for Systems 2, 3 and 4.

System 2	Sys. task completion	Q. task completion	Call duration	Number of ASR rejects	Q. ASR rating	Q. phrasing rating	Q. TTS rating	Q. overall rating	Q. task ease	Q. future use
Q. task completion	0,66 0,56 0,73 0,66									
	0,59 0,59 0,54 0,66									
	0,68 0,68 0,83 0,57									
	0,69 0,35 0,78 0,68									
Call duration	-0,08 -0,18 -0,10 0,05 -0,24 -0,33 -0,22 -0,16									
	-0,07 -0,17 -0,06 0,19 -0,23 -0,37 -0,18 0,03									
	-0,08 -0,27 -0,05 -0,02 -0,19 -0,28 -0,04 -0,21									
	-0,07 -0,09 -0,11 0,12 -0,29 -0,28 -0,33 -0,18									
Number of ASR rejects	NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN									
	NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN									
	NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN									
	NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN									
Q. ASR rating	0,31 0,30 0,27 0,36 0,45 0,46 0,44 0,43 -0,35 -0,41 -0,35 -0,25 NaN NaN NaN NaN									
	0,28 0,28 0,18 0,45 0,43 0,46 0,53 0,25 -0,33 -0,33 -0,36 -0,33 NaN NaN NaN NaN									
	0,28 0,44 0,35 0,13 0,43 0,59 0,41 0,37 -0,31 -0,56 -0,04 -0,31 NaN NaN NaN NaN									
	0,36 0,14 0,28 0,45 0,46 0,36 0,41 0,51 -0,39 -0,46 -0,48 -0,14 NaN NaN NaN NaN									
Q. phrasing rating	0,08 0,06 0,09 0,09 0,15 0,10 0,18 0,15 -0,11 -0,14 -0,08 -0,11 NaN NaN NaN NaN 0,34 0,29 0,46 0,26									
	0,14 0,10 0,07 0,35 0,22 0,16 0,33 0,18 -0,10 -0,11 -0,04 -0,19 NaN NaN NaN NaN 0,40 0,29 0,47 0,51									
	-0,02 -0,12 0,12 -0,07 0,16 0,11 0,16 0,19 -0,09 -0,16 0,12 -0,16 NaN NaN NaN NaN 0,33 0,24 0,43 0,33									
	0,10 0,19 0,07 0,09 0,07 0,01 0,11 0,04 -0,12 -0,21 -0,16 0,01 NaN NaN NaN NaN 0,29 0,36 0,47 -0,02									
Q. TTS rating	0,11 0,12 0,00 0,24 0,14 0,18 0,11 0,14 -0,06 -0,13 -0,06 -0,01 NaN NaN NaN NaN 0,35 0,37 0,26 0,42 0,34 0,35 0,26 0,42									
	0,06 0,11 -0,15 0,26 0,11 0,26 0,00 -0,07 -0,04 -0,06 -0,02 0,00 NaN NaN NaN NaN 0,36 0,43 0,18 0,52 0,31 0,36 0,14 0,47									
	0,06 0,08 -0,01 0,12 0,16 0,17 0,13 0,19 -0,08 -0,34 0,38 -0,13 NaN NaN NaN NaN 0,33 0,21 0,19 0,49 0,44 0,38 0,56 0,40									
	0,20 0,20 0,10 0,37 0,13 0,02 0,16 0,17 -0,08 -0,17 -0,25 0,19 NaN NaN NaN NaN 0,34 0,43 0,39 0,26 0,29 0,34 0,18 0,41									
Q. overall rating	0,46 0,44 0,45 0,46 0,57 0,56 0,58 0,57 -0,34 -0,44 -0,28 -0,24 NaN NaN NaN NaN 0,66 0,69 0,65 0,64 0,37 0,36 0,41 0,35 0,33 0,36 0,25 0,39									
	0,42 0,42 0,37 0,50 0,57 0,55 0,66 0,47 -0,35 -0,41 -0,29 -0,22 NaN NaN NaN NaN 0,70 0,70 0,78 0,61 0,37 0,31 0,46 0,38 0,36 0,38 0,24 0,49									
	0,39 0,49 0,49 0,28 0,53 0,60 0,58 0,47 -0,26 -0,48 -0,01 -0,24 NaN NaN NaN NaN 0,62 0,65 0,60 0,63 0,38 0,28 0,44 0,42 0,30 0,17 0,14 0,43									
	0,53 0,43 0,48 0,56 0,61 0,55 0,54 0,68 -0,37 -0,45 -0,35 -0,26 NaN NaN NaN NaN 0,65 0,72 0,59 0,62 0,36 0,55 0,37 0,24 0,32 0,50 0,32 0,25									
Q. task ease	0,30 0,25 0,38 0,24 0,64 0,66 0,66 0,59 -0,37 -0,36 -0,37 -0,31 NaN NaN NaN NaN 0,68 0,67 0,71 0,65 0,28 0,30 0,29 0,28 0,24 0,29 0,22 0,23 0,64 0,62 0,64 0,63									
	0,33 0,34 0,27 0,46 0,70 0,70 0,74 0,70 -0,28 -0,26 -0,28 -0,35 NaN NaN NaN NaN 0,64 0,66 0,71 0,49 0,36 0,32 0,41 0,36 0,25 0,41 0,07 0,15 0,65 0,61 0,76 0,62									
	0,28 0,26 0,56 0,09 0,59 0,57 0,71 0,52 -0,35 -0,51 -0,14 -0,37 NaN NaN NaN NaN 0,73 0,78 0,74 0,69 0,34 0,31 0,27 0,43 0,28 0,18 0,22 0,38 0,63 0,63 0,62 0,64									
	0,28 0,02 0,35 0,20 0,61 0,67 0,58 0,57 -0,45 -0,45 -0,53 -0,24 NaN NaN NaN NaN 0,68 0,61 0,68 0,68 0,17 0,26 0,21 0,05 0,20 0,15 0,35 0,08 0,62 0,64 0,57 0,59									
Q. future use	0,26 0,18 0,33 0,24 0,59 0,59 0,60 0,58 -0,29 -0,27 -0,29 -0,28 NaN NaN NaN NaN 0,64 0,64 0,66 0,60 0,28 0,29 0,28 0,27 0,30 0,38 0,24 0,30 0,60 0,58 0,61 0,60 0,88 0,87 0,89 0,87									
	0,28 0,28 0,20 0,46 0,64 0,62 0,63 0,70 -0,20 -0,19 -0,21 -0,23 NaN NaN NaN NaN 0,63 0,67 0,70 0,41 0,35 0,32 0,38 0,38 0,29 0,45 0,15 0,16 0,60 0,55 0,72 0,55 0,89 0,88 0,90 0,88									
	0,28 0,18 0,52 0,15 0,56 0,53 0,65 0,52 -0,26 -0,37 -0,13 -0,28 NaN NaN NaN NaN 0,65 0,67 0,69 0,62 0,30 0,22 0,28 0,36 0,33 0,33 0,16 0,46 0,61 0,56 0,65 0,63 0,85 0,83 0,86 0,86									
	0,23 -0,08 0,30 0,16 0,58 0,59 0,55 0,54 -0,40 -0,37 -0,43 -0,33 NaN NaN NaN NaN 0,62 0,56 0,60 0,65 0,19 0,31 0,21 0,06 0,27 0,27 0,36 0,15 0,58 0,67 0,51 0,55 0,89 0,90 0,89 0,88									
Q. number of call	0,04 0,06 -0,02 0,05 0,01 -0,02 0,03 -0,01 -0,16 -0,15 -0,23 -0,07 NaN NaN NaN NaN 0,08 0,08 0,04 0,10 -0,01 0,13 -0,09 -0,05 0,01 -0,08 0,09 0,02 0,04 0,12 0,02 -0,05 0,04 -0,04 0,09 0,00 0,02 -0,06 0,09 0,00									
	-0,01 -0,03 0,00 0,09 -0,11 -0,11 -0,12 -0,14 -0,09 -0,11 -0,15 0,06 NaN NaN NaN NaN 0,00 0,03 -0,14 0,14 -0,05 0,12 -0,24 -0,15 -0,10 -0,16 0,01 -0,11 -0,01 0,08 0,00 -0,29 -0,07 -0,11 -0,04 -0,13 -0,04 -0,09 0,04 -0,09									
	0,10 0,24 0,02 0,10 0,01 0,14 -0,03 -0,04 -0,15 -0,18 -0,30 -0,05 NaN NaN NaN NaN 0,12 0,11 0,14 0,11 0,02 0,14 0,05 -0,04 0,06 0,08 0,03 0,06 0,03 0,22 -0,09 -0,01 0,03 -0,02 0,15 -0,03 0,02 0,00 0,12 -0,02									
	0,03 0,15 -0,03 -0,01 0,11 0,03 0,14 0,11 -0,25 -0,30 -0,28 -0,17 NaN NaN NaN NaN 0,13 0,19 0,12 0,06 0,00 0,14 -0,07 -0,01 0,08 -0,02 0,16 0,06 0,10 0,14 0,09 0,04 0,15 0,12 0,15 0,14 0,08 -0,03 0,13 0,08									

Figure 3: Systems evaluation: Correlations between key performance indicators (see section 3.3 for details) for System 2.

System 3	Sys. task completion	Q. task completion	Call duration	Number of ASR rejects	Q. ASR rating	Q. phrasing rating	Q. TTS rating	Q. overall rating	Q. task ease	Q. future use
Q. task completion	0,88 0,97 0,82 0,85									
	0,91 0,97 0,84 0,87									
	0,85 0,96 0,78 0,81									
	0,89 1,00 0,83 0,90									
Call duration	-0,47 -0,89 -0,34 -0,05	-0,50 -0,89 -0,41 -0,12								
	-0,67 -0,90 -0,51 -0,42	-0,72 -0,91 -0,57 -0,57								
	-0,28 -0,87 0,54 0,31	-0,30 -0,84 0,58 0,23								
	-0,44 -0,94 -0,35 -0,18	-0,48 -0,94 -0,44 -0,20								
Number of ASR rejects	-0,25 -0,43 -0,20 -0,18	-0,23 -0,43 -0,21 -0,14	0,40 0,51 0,46 0,32							
	-0,29 -0,43 -0,24 -0,18	-0,32 -0,45 -0,26 -0,25	0,46 0,55 0,43 0,39							
	-0,16 -0,56 0,13 -0,01	-0,12 -0,53 0,13 0,08	0,37 0,50 0,68 0,27							
	-0,32 -0,36 -0,31 -0,31	-0,29 -0,36 -0,33 -0,26	0,45 0,57 0,59 0,35							
Q. ASR rating	0,47 0,62 0,34 0,43	0,50 0,60 0,39 0,51	-0,29 -0,51 -0,12 -0,07	-0,32 -0,39 -0,38 -0,24						
	0,46 0,63 0,33 0,33	0,49 0,59 0,40 0,43	-0,30 -0,49 -0,07 -0,13	-0,27 -0,35 -0,34 -0,10						
	0,51 0,71 0,42 0,41	0,55 0,70 0,48 0,47	-0,24 -0,56 0,04 0,14	-0,36 -0,59 -0,45 -0,19						
	0,42 0,42 0,32 0,49	0,47 0,42 0,38 0,57	-0,31 -0,42 -0,21 -0,17	-0,32 -0,18 -0,37 -0,29						
Q. phrasing rating	0,09 0,08 0,18 0,03	0,13 0,08 0,22 0,09	-0,12 -0,04 -0,17 -0,10	-0,09 -0,03 -0,11 -0,07	0,27 0,18 0,32 0,25					
	0,19 0,17 0,42 -0,01	0,24 0,18 0,51 0,04	-0,13 -0,17 -0,11 -0,05	-0,28 -0,17 -0,35 -0,34	0,34 0,25 0,39 0,42					
	-0,02 0,03 -0,01 -0,07	0,04 0,02 0,14 0,03	-0,09 0,09 -0,05 -0,14	0,04 0,06 0,05 0,04	0,12 0,13 0,12 0,09					
	0,11 -0,08 0,07 0,16	0,11 -0,08 0,05 0,19	-0,14 0,09 -0,25 -0,06	-0,09 -0,03 -0,07 -0,06	0,36 0,20 0,33 0,35					
Q. TTS rating	0,10 0,19 0,08 0,02	0,11 0,19 0,09 0,05	-0,10 -0,13 -0,07 -0,06	-0,06 -0,06 -0,13 -0,03	0,24 0,28 0,24 0,20	0,35 0,38 0,45 0,26				
	0,12 0,18 0,10 -0,01	0,18 0,19 0,27 0,07	-0,12 -0,18 -0,02 -0,21	-0,20 -0,28 -0,19 -0,15	0,28 0,26 0,25 0,36	0,38 0,38 0,58 0,21				
	0,16 0,33 0,15 0,03	0,15 0,33 0,15 0,03	-0,09 -0,16 0,16 0,01	0,01 -0,05 -0,03 0,07	0,28 0,48 0,21 0,15	0,27 0,46 0,20 0,22				
	0,04 0,05 0,06 -0,01	0,02 0,05 -0,01 0,03	-0,09 -0,04 -0,16 0,00	-0,05 0,07 -0,15 -0,04	0,18 0,09 0,24 0,12	0,37 0,33 0,44 0,34				
Q. overall rating	0,53 0,63 0,46 0,48	0,59 0,65 0,53 0,59	-0,37 -0,57 -0,15 -0,19	-0,34 -0,43 -0,34 -0,28	0,75 0,71 0,75 0,77	0,30 0,25 0,35 0,28	0,39 0,43 0,41 0,34			
	0,50 0,56 0,43 0,45	0,59 0,59 0,62 0,57	-0,34 -0,49 -0,12 -0,19	-0,29 -0,40 -0,31 -0,14	0,76 0,73 0,75 0,84	0,40 0,33 0,61 0,34	0,49 0,48 0,59 0,44			
	0,49 0,71 0,32 0,35	0,56 0,73 0,32 0,51	-0,33 -0,57 -0,06 0,02	-0,32 -0,48 -0,44 -0,16	0,76 0,79 0,75 0,73	0,17 0,22 0,15 0,14	0,39 0,53 0,27 0,29			
	0,58 0,63 0,52 0,60	0,62 0,63 0,55 0,67	-0,42 -0,68 -0,22 -0,28	-0,38 -0,47 -0,33 -0,36	0,73 0,52 0,76 0,75	0,33 0,17 0,27 0,36	0,32 0,26 0,36 0,30			
Q. task ease	0,66 0,87 0,57 0,50	0,73 0,89 0,68 0,61	-0,70 -0,87 -0,57 -0,50	-0,38 -0,50 -0,39 -0,29	0,60 0,68 0,56 0,53	0,26 0,20 0,36 0,22	0,24 0,24 0,26 0,21	0,69 0,69 0,66 0,68		
	0,81 0,91 0,77 0,61	0,86 0,93 0,83 0,75	-0,82 -0,92 -0,70 -0,79	-0,40 -0,51 -0,33 -0,35	0,57 0,66 0,45 0,53	0,35 0,32 0,49 0,27	0,27 0,26 0,23 0,37	0,62 0,65 0,55 0,61		
	0,55 0,84 0,22 0,38	0,66 0,89 0,52 0,51	-0,59 -0,83 0,01 -0,27	-0,30 -0,58 -0,30 -0,08	0,58 0,73 0,67 0,42	0,20 -0,01 0,29 0,25	0,21 0,29 0,17 0,15	0,67 0,71 0,68 0,62		
	0,62 0,83 0,55 0,57	0,67 0,83 0,63 0,64	-0,66 -0,87 -0,58 -0,50	-0,46 -0,51 -0,48 -0,43	0,65 0,60 0,60 0,64	0,22 0,11 0,27 0,14	0,23 0,17 0,31 0,14	0,77 0,78 0,73 0,75		
Q. future use	0,65 0,77 0,56 0,59	0,71 0,80 0,62 0,69	-0,59 -0,78 -0,47 -0,38	-0,34 -0,45 -0,34 -0,27	0,60 0,63 0,55 0,60	0,29 0,24 0,37 0,27	0,31 0,29 0,35 0,31	0,75 0,76 0,70 0,76	0,88 0,89 0,87 0,87	
	0,74 0,81 0,70 0,63	0,81 0,83 0,80 0,79	-0,73 -0,82 -0,66 -0,66	-0,36 -0,47 -0,27 -0,32	0,57 0,63 0,42 0,62	0,35 0,34 0,49 0,22	0,37 0,34 0,34 0,48	0,69 0,71 0,57 0,75	0,91 0,91 0,91 0,89	
	0,60 0,80 0,45 0,48	0,70 0,84 0,64 0,60	-0,50 -0,81 0,19 -0,11	-0,29 -0,55 -0,23 -0,11	0,61 0,70 0,70 0,50	0,22 0,11 0,29 0,25	0,28 0,34 0,21 0,24	0,73 0,79 0,66 0,68	0,86 0,88 0,86 0,83	
	0,61 0,63 0,52 0,69	0,62 0,63 0,52 0,73	-0,51 -0,66 -0,45 -0,39	-0,39 -0,41 -0,42 -0,36	0,62 0,44 0,58 0,67	0,31 0,15 0,32 0,30	0,30 0,13 0,40 0,25	0,81 0,81 0,78 0,82	0,86 0,86 0,81 0,85	0,87
Q. number of call	0,04 0,10 -0,01 0,03	0,00 0,09 -0,08 -0,02	-0,06 -0,12 0,06 -0,07	-0,09 0,00 -0,01 -0,19	-0,03 -0,03 -0,12 0,03	-0,07 -0,06 -0,10 -0,08	0,04 0,07 0,01 0,03	0,02 0,08 -0,09 0,04	-0,03 0,03 -0,17 0,00	0,03 0,12 -0,08 0,02
	-0,03 0,03 -0,15 0,00	-0,04 0,00 -0,16 0,04	-0,04 -0,05 0,13 -0,15	-0,05 -0,05 0,12 -0,20	-0,07 -0,12 -0,13 0,08	-0,03 -0,03 -0,13 0,06	0,03 0,10 -0,12 0,05	-0,02 -0,03 -0,14 0,10	-0,04 -0,01 -0,21 0,08	-0,02 -0,02 -0,01 -0,19 0,11
	0,03 0,10 0,00 0,00	-0,02 0,12 -0,17 -0,05	-0,01 -0,11 -0,12 0,17	-0,04 0,14 -0,07 -0,15	0,02 0,09 -0,02 -0,01	-0,09 0,12 -0,31 -0,12	0,09 0,23 0,15 -0,04	0,00 0,18 -0,09 -0,12	-0,13 0,05 -0,21 -0,28	-0,03 0,23 -0,14 -0,21
	0,10 0,25 0,07 0,06	0,05 0,25 0,01 -0,02	-0,13 -0,24 0,02 -0,26	-0,14 -0,11 -0,03 -0,27	-0,06 -0,04 -0,15 0,05	-0,10 -0,27 0,00 -0,13	0,01 -0,12 0,03 0,10	0,06 0,17 -0,07 0,18	0,06 0,13 -0,12 0,28	0,12 0,27 0,00 0,23

Figure 4: Systems evaluation: Correlations between key performance indicators (see section 3.3 for details) for System 3.

System 4	Sys. task completion	Q. task completion	Call duration	Number of ASR rejects	Q. ASR rating	Q. phrasing rating	Q. TTS rating	Q. overall rating	Q. task ease	Q. future use
Q. task completion	0,83 0,75 0,89 0,85									
	0,81 0,75 1,00 0,84									
	0,89 0,80 NaN 0,92									
	0,80 0,72 0,86 0,70									
Call duration	-0,23 -0,16 -0,34 -0,20	-0,32 -0,29 -0,38 -0,30								
	-0,32 -0,30 -0,55 -0,18	-0,40 -0,40 -0,55 -0,28								
	-0,14 0,03 NaN -0,20	-0,25 -0,17 NaN -0,28								
	-0,17 0,22 -0,22 -0,14	-0,27 0,14 -0,29 -0,29								
Number of ASR rejects	-0,29 -0,21 -0,39 -0,33	-0,35 -0,30 -0,41 -0,38	0,65 0,65 0,67 0,65							
	-0,25 -0,24 -0,25 -0,30	-0,31 -0,33 -0,25 -0,30	0,64 0,65 0,66 0,59							
	-0,31 -0,17 NaN -0,37	-0,38 -0,28 NaN -0,43	0,61 0,66 -0,28 0,59							
	-0,33 -0,19 -0,37 -0,26	-0,38 -0,21 -0,40 -0,33	0,68 0,52 0,65 0,78							
Q. ASR rating	0,30 0,26 0,30 0,33	0,37 0,39 0,30 0,39	-0,40 -0,36 -0,33 -0,52	-0,44 -0,46 -0,39 -0,48						
	0,40 0,36 0,31 0,50	0,48 0,48 0,31 0,57	-0,51 -0,42 -0,50 -0,75	-0,51 -0,52 -0,55 -0,62						
	0,28 0,30 NaN 0,27	0,32 0,34 NaN 0,30	-0,33 -0,25 0,28 -0,35	-0,44 -0,39 -1,00 -0,45						
	0,21 0,01 0,24 0,25	0,29 0,21 0,25 0,36	-0,30 -0,32 -0,18 -0,46	-0,36 -0,45 -0,29 -0,40						
Q. phrasing rating	0,08 0,06 0,15 0,04	0,17 0,20 0,17 0,13	-0,13 -0,15 -0,06 -0,16	-0,13 -0,08 -0,16 -0,18	0,29 0,36 0,22 0,29					
	0,03 0,05 0,15 -0,08	0,20 0,30 0,15 0,00	-0,16 -0,14 -0,13 -0,28	-0,13 -0,08 -0,17 -0,46	0,31 0,28 0,41 0,31					
	0,11 0,18 NaN 0,08	0,13 0,17 NaN 0,12	-0,14 -0,23 0,28 -0,11	-0,13 -0,10 -1,00 -0,14	0,32 0,52 1,00 0,22					
	0,07 -0,09 0,14 -0,01	0,16 -0,08 0,16 0,20	-0,09 -0,18 0,00 -0,12	-0,12 -0,02 -0,13 -0,09	0,25 0,33 0,15 0,35					
Q. TTS rating	0,12 0,06 0,17 0,12	0,14 0,14 0,12 0,15	-0,11 0,01 -0,14 -0,21	-0,10 0,00 -0,14 -0,21	0,36 0,34 0,31 0,40	0,31 0,29 0,28 0,33				
	0,09 0,10 -0,06 0,20	0,12 0,16 -0,06 0,20	-0,03 0,01 -0,01 -0,26	-0,02 0,00 -0,11 -0,18	0,32 0,29 0,27 0,44	0,35 0,28 0,52 0,33				
	0,12 0,07 NaN 0,13	0,13 0,12 NaN 0,12	-0,08 -0,10 0,28 -0,08	-0,17 -0,14 -1,00 -0,18	0,42 0,46 1,00 0,41	0,27 0,26 1,00 0,27				
	0,13 -0,02 0,22 0,05	0,16 0,12 0,16 0,15	-0,20 0,09 -0,17 -0,41	-0,14 0,18 -0,13 -0,29	0,35 0,36 0,31 0,39	0,30 0,37 0,17 0,44				
Q. overall rating	0,42 0,33 0,45 0,47	0,51 0,51 0,46 0,54	-0,46 -0,37 -0,51 -0,53	-0,42 -0,40 -0,42 -0,46	0,75 0,74 0,70 0,80	0,33 0,36 0,32 0,31	0,45 0,37 0,42 0,53			
	0,43 0,37 0,40 0,57	0,52 0,51 0,40 0,59	-0,49 -0,43 -0,56 -0,65	-0,46 -0,48 -0,49 -0,53	0,81 0,81 0,71 0,88	0,38 0,39 0,41 0,34	0,44 0,37 0,50 0,57			
	0,46 0,50 NaN 0,44	0,53 0,65 NaN 0,49	-0,34 -0,25 0,72 -0,36	-0,36 -0,28 -0,87 -0,39	0,70 0,59 0,87 0,74	0,29 0,33 0,87 0,26	0,46 0,30 0,87 0,52			
	0,36 0,10 0,42 0,42	0,47 0,35 0,44 0,56	-0,48 -0,38 -0,43 -0,62	-0,41 -0,35 -0,35 -0,51	0,72 0,72 0,66 0,82	0,31 0,32 0,27 0,34	0,43 0,42 0,37 0,55			
Q. task ease	0,54 0,47 0,65 0,50	0,64 0,65 0,67 0,60	-0,63 -0,53 -0,69 -0,69	-0,48 -0,45 -0,55 -0,50	0,66 0,66 0,61 0,71	0,27 0,34 0,23 0,24	0,27 0,23 0,23 0,33	0,77 0,77 0,77 0,78		
	0,56 0,52 0,70 0,57	0,66 0,66 0,70 0,63	-0,67 -0,62 -0,78 -0,72	-0,49 -0,51 -0,47 -0,54	0,73 0,66 0,67 0,91	0,32 0,35 0,27 0,31	0,22 0,17 0,09 0,50	0,76 0,70 0,74 0,91		
	0,54 0,62 NaN 0,51	0,64 0,78 NaN 0,59	-0,57 -0,30 0,28 -0,67	-0,46 -0,33 -1,00 -0,51	0,63 0,61 1,00 0,64	0,26 0,36 1,00 0,20	0,26 0,29 1,00 0,25	0,75 0,85 0,87 0,70		
	0,51 0,22 0,62 0,41	0,62 0,44 0,65 0,60	-0,61 -0,46 -0,60 -0,62	-0,50 -0,43 -0,52 -0,43	0,60 0,73 0,53 0,58	0,24 0,28 0,20 0,24	0,31 0,34 0,27 0,35	0,79 0,89 0,75 0,78		
Q. future use	0,53 0,45 0,65 0,49	0,65 0,65 0,70 0,60	-0,50 -0,37 -0,58 -0,58	-0,39 -0,34 -0,46 -0,44	0,62 0,66 0,56 0,63	0,28 0,40 0,22 0,23	0,37 0,36 0,31 0,42	0,79 0,83 0,75 0,80	0,86 0,84 0,88 0,85	
	0,53 0,49 0,67 0,53	0,67 0,68 0,67 0,65	-0,51 -0,43 -0,66 -0,60	-0,38 -0,39 -0,42 -0,38	0,72 0,73 0,63 0,79	0,36 0,41 0,30 0,30	0,36 0,34 0,24 0,54	0,83 0,82 0,79 0,88	0,84 0,81 0,87 0,88	
	0,52 0,60 NaN 0,49	0,64 0,78 NaN 0,58	-0,42 -0,23 0,28 -0,49	-0,37 -0,28 -1,00 -0,41	0,59 0,55 1,00 0,60	0,29 0,38 1,00 0,24	0,35 0,33 1,00 0,36	0,80 0,83 0,87 0,78	0,86 0,89 1,00 0,85	
	0,51 0,20 0,62 0,42	0,63 0,39 0,69 0,59	-0,53 -0,31 -0,50 -0,63	-0,43 -0,20 -0,42 -0,49	0,53 0,60 0,49 0,49	0,21 0,37 0,18 0,13	0,39 0,46 0,32 0,47	0,75 0,88 0,71 0,71	0,86 0,86 0,86 0,82	
Q. number of call	0,03 0,05 0,03 0,00	0,01 0,04 0,05 -0,05	0,03 0,06 0,03 0,02	-0,07 -0,06 -0,03 -0,13	0,06 0,10 0,03 0,05	-0,04 0,01 -0,13 -0,01	0,06 0,08 0,06 0,03	0,03 0,06 0,00 0,02	0,00 -0,03 0,02 0,01	0,04 0,02 0,06 0,01
	-0,01 -0,04 -0,08 0,14	-0,03 -0,01 -0,08 -0,01	0,12 0,14 0,11 0,07	-0,03 0,02 -0,05 -0,27	0,04 0,02 0,06 0,10	-0,01 -0,02 0,16 -0,10	-0,01 0,10 -0,10 -0,12	0,00 -0,02 -0,03 0,10	-0,09 -0,17 -0,04 0,07	-0,05 -0,06 -0,09 0,01
	0,01 0,17 NaN -0,06	0,01 0,14 NaN -0,05	-0,01 -0,16 -0,84 0,06	-0,15 -0,36 0,76 -0,05	0,12 0,23 -0,76 0,07	0,06 0,08 -0,76 0,07	0,09 -0,05 -0,76 0,16	0,06 0,15 -0,98 0,03	0,05 0,15 -0,76 0,00	0,05 0,10 -0,76 0,03
	0,07 0,12 0,06 0,01	0,04 0,03 0,09 -0,11	-0,04 0,00 -0,01 -0,11	-0,08 -0,08 -0,03 -0,18	0,04 0,16 0,02 -0,05	-0,14 0,00 -0,21 -0,12	0,08 0,14 0,14 -0,12	0,04 0,15 0,03 -0,12	0,05 0,14 0,05 -0,06	0,10 0,13 0,13 -0,03

Figure 5: Systems evaluation: Correlations between key performance indicators (see section 3.3 for details) for System 4.

Appendix B

Transcription Conventions (Conversation Analysis)

B.1 Temporal and sequential relationships

Sign	Meaning
[ ]	Brackets bridging two utterances by different speakers indicate overlap. Left brackets for the beginning of overlap and right ones for its end.
=	Equal signs (1st sign at the end of a line and second sign at the beginning of another line) indicate that there is no discernible interval between 1st and 2nd speakers' turns.
&	Ampersands (1st sign at the end of a line and second sign at the beginning of another line). Indicates the continuation of turn for a same speaker.
(#)	A pause timed in seconds. # is a number.
(.) (..) (...)	A dot in parentheses indicates a noticeable pause too short to measure.

B.2 Aspects of Speech Delivery and Intonation

Sign	Meaning
.	Periods indicate a falling intonation, not necessarily the end of a sentence.
?	Questions marks indicate rising intonation, not necessarily a question.
,	Commas indicate "continuing" intonation.
::	Colons indicate the lengthening of a vowel; The more colons, the longer the lengthening.
-	Hyphens after a word or part of it indicate a cut-off or self-interruption
word	Underlining indicates stress or emphasis.
WOrd	Louder talk is transcribed in capital letters.
° °	Talk between degree signs is markedly quiet or soft.

Sign	Meaning
$>\ <$	Talk between $>$ and $<$ signs is compressed or rushed.
$<\ >$	In the reverse order, it is markedly slow or drawn out.
hh	The letter h indicates an audible out-breathing - the more hs, the more aspiration.
hh$^\circ$	Audible in-breathing.

B.3 Other Markings

Sign	Meaning
(word)	All or part of an utterance in parentheses indicates uncertainty on the transcriber's part, but represents a likely possibility.
xxx	Inaudible segment.
(( ))	Description of non-transcribed events is between double parentheses.