Examining the Effects of Rephrasing User Input on Two Mobile Spoken Language Systems

Nikos Tsourakis[1], Agnes Lisowska[2], Manny Rayner[1], Pierrette Bouillon[1]
[1]ISSCO/TIM/ETI, University of Geneva, Switzerland
Boulevard du Pont-d'Arve, CH-1211 Genève 4, Switzerland
[2]Pervasive and Artificial Intelligence Group, University of Fribourg, Switzerland
Boulevard de Pérolles 90, CH-1700, Fribourg, Switzerland
E-mail: {Nikolaos.Tsourakis, Emmanuel.Rayner, Pierrette.Bouillon}@unige.ch, Agnes.Lisowska@unifr.ch

Abstract

In this work we investigate the effects of rephrasing the user's input on two mobile spoken dialogue systems. We argue that for specific kinds of applications it's important to confirm the understanding of the system before obtaining the output. In this way the user can avoid misconceptions and problems occurring in the dialogue flow and he can enhance his confidence in the system. Nevertheless this has an impact on the interaction, as the mental workload increases, and the user's behavior may adapt to the system's coverage. We will focus on two applications that implement the notion of rephrasing user's input in a different way. Our study took place among 14 subjects that used both systems on a Nokia N810 Internet Tablet.

1. Introduction

During the construction of a spoken dialogue system much effort is spent on improving the quality of speech recognition as possible. However, even if an application perfectly recognizes the input, its understanding may be far from what the user originally meant. Consider for example a request like "When is the meeting next Friday?". It can be equally interpreted as "When is the meeting on the closest Friday to today?" or "When is the meeting on Friday next week?". The user should be informed about what the system actually understood so that an error will not have a negative impact in the later stages of the dialogue (Walker et al, 1998), and the user will not perceive a bad response as correct and vice-versa, leading to an increase in their confusion and cognitive load (Weegels, 2004).

One important aspect that this work tries to address is the effect of presenting the system's understanding during interaction with users. This is actually an enriched version of the user's input, taking into account constraints of the application, the dialogue's current situation etc. We investigate the following issues:

- What is the mental effort considering the fact that the user takes some time to understand the output?
- Does the rephrased output direct users to the coverage of the application (Zoltan & Ford, 1991)?
- Does the psychological notion of free recall of text, where adults are likely to reproduce the gist, or "essence" of a text instead of its verbatim reproduction apply in our situation (Clark & Clark, 1977)?
- To what extent does a user confirm wrong output and discard correct output.
- How does the user behave after long successful or failed interactions? Do they confirm the output with less thought?
- Is rephrasing a good way to hide recognition errors?

Previous studies focused on the influence of system output style (personal/impersonal) for the users' subjective judgments of a system (Nass & Brave, 2005), as well as their formulation of input (Brennan & Ohaeri, 1994). Other studies examine how to enhance beliefs in the system's output (Bohus & Rudnicky, 2005). Our work focuses more on the explanatory structure of this output. We therefore implemented two applications, based on the Regulus Open Source platform (Rayner et al, 2006), where the notion of rephrasing user input has been put into place.

The first system is a Calendar application (Tsourakis et al, 2008) for accessing past and future meetings, along with information about the participants. In this context we introduce a rephrasing mechanism called paraphrases, which further analyzes user input and presents its enriched representation by considering different dialogue constraints.

The second system is a medical speech translator (Bouillon et al, 2005), MedSLT, with which doctors can ask foreign patients medical diagnosis questions. These questions are translated and announced in the patient's language. In order to confirm the system's understanding the doctor is shown the back-translation of his input (e.g. translation from English-to-English).

This paper is organized as follows. In section 2 we give a short overview of the Regulus platform followed by a presentation of specific features of the platform related to the current work. We continue in Section 3 with a description of the experimental set-up for both systems. In Section 4 the results of the evaluation with real users are presented along with a discussion. We conclude in the final section.

2. Regulus

Regulus (Rayner et al, 2006) is an Open Source toolkit for building medium vocabulary grammar-based spoken dialogue and translation systems. The central idea is to base run-time processing on efficient, task-specific grammars derived from general, reusable, domain-independent core grammars. A detailed description of the core grammar for English can be found in Chapter 9 of the book.

The core grammars are automatically specialized, using corpus-driven methods based on small corpora, to derive simpler grammars. Specialization is both with respect to task (recognition, analysis, generation) and to application domain. Each of these specialized unification grammars is then subjected to a second compilation step, which converts it into its executable form. For analysis and generation, this form is a standard parser or generator. For recognition, it is a semantically annotated CFG grammar in the form required by the Nuance engine, which is then subjected to further Nuance-specific compilation steps to derive a speech recognition package.

The Regulus platform also contains further infrastructure to support construction of applications which use the recognizers, parsers and generators as components. In both cases, the main processing flow consists of a pipeline. Thus processing in a speech translation application starts with speech recognition (including parsing), which produces a source language semantic representation. This representation is then passed to a translation engine, which first translates it into an interlingua form, and then into a target language representation. Finally, the target language grammar, compiled into generation form, is used to create a target language surface string.

The generic dialogue application architecture is similar. The central component is the Dialogue Manager (DM), which receives dialogue moves and produces abstract actions. It also manipulates an information state, which maintains context; processing will generally be context-dependent. The DM is bracketed between two other components, the Input Manager (IM) and the Output Manager (OM). The IM receives logical forms, and non-speech inputs if there are any, and turns them into dialogue moves. The OM received abstract actions and turns them into concrete actions. Usually, these actions will be either speaking, though TTS or recorded speech, or manipulation of a GUI's screen area. The speech translation and dialogue application architectures are described in detail in Chapters 5 and 6 of (Rayner et al, 2006).

2.1 Rephrasing user's input

As noted in Section 1, we deem important to provide the users with some indication of how a speech translation or spoken dialogue system has interpreted their input. We have experimented with slightly different mechanisms in the two cases. For speech translation, we perform a "back-translation" from the interlingua; we apply rules to translate from the interlingua to the source language, and then use a generator derived from the source language to produce a source language surface string. Given that the system is already capable of multi-lingual translation, and in particular of realizing an interlingua form in the different languages, this strategy is very easy to realize.

No corresponding mechanisms exist in the case of a dialogue application, where the level of representation corresponding to the interlingua is the dialogue move. The solution we have instead chosen is to implement a grammar, again compiled into a generator, which associates a surface string with each dialogue move. So far, we have had two main design goals for these "paraphrase grammars"; the surface form for the paraphrase should be unambiguous, and it should also be fairly natural. These two goals conflict to some extent, since completely natural language is typically ambiguous to some degree. For the applications we have so far been involved with, it has however proved feasible to find a reasonable tradeoff point; the paraphrase grammars can also be kept simple and compact. Table 1 shows some examples for both cases.

User:	**Paraphrase:**
"Is there a meeting on Friday morning?"	Is there a meeting between 06:00 and 12:00 on Fri Mar 19 2010?
"Was there a meeting last week?"	What meetings were there between Mon Mar 8 2010 and Sun Mar 15 2010?
"Do I have a meeting tomorrow?"	Are there meetings on Wed Apr 17 2010 attended by Nikos Tsourakis?
"Is someone from Geneva coming?"	Who is attending the meeting affiliated with Geneva?
User:	**Back-translation:**
"Does red wine cause any headaches?"	Do you have the headaches when you drink red wine?
"What relieves your pain?"	What makes the pain better?
"How about bright light?"	Is your headache made worse by bright light?
"Do you have it every day?"	Does the pain occur every day?

Table 1: Examples of paraphrases/back-translation

3. Experimental Design

In order to examine our ideas we set up a sequence of experiments using the two systems. It is important to remark that we don't try to compare the two applications as they are different in nature. Instead we seek to extract uniform results, as both implement the notion of rephrasing user's input in different ways. For the evaluation we used the GUI presented in Figure 1, running on the Nokia's N810 Internet Tablet.

Figure 1: Evaluation GUI

14 participants were split into two groups of 7 subjects. The first group used the Calendar application whereas the second used the MedSLT system. Each user in the first group was given a set of 30 scenarios that demanded just one interaction. The idea was simple: they had to confirm

whether the system could understand what they asked for. The description of the first 15 scenarios (first session) was well defined and specific, e.g. "You have a meeting with Alex, but you don't remember the time". The last 15 scenarios (second session) were more ambiguous and the users could pose a question in different ways. For example for the scenario "Meeting in September?" a user could say "Was there a meeting in September?" or "When is the next meeting in September?". In this way we avoid imposing on the user what to say and conversely investigate possible learning effects from the previous interactions. Some of these scenarios contained images of persons, places etc (Figure 2) in order to let the users improvise. Between the two sessions there was a short break of 5 minutes.



When?

Figure 2: Image scenario (When is the next meeting with Brad Pitt in Geneva?)

Our experiments were organized into two configurations. In the first one the system responded to the user's input with the recognition result and in the second configuration with the paraphrase. These two versions were evaluated on different dates and the group consisted of non native, fluent English speakers, with proficient computer skills.

After speaking to the system the user had to confirm the output. A positive confirmation ("Yes" button) was given if the output expressed the same semantic meaning as the input, otherwise the "No" button was pressed. The turns per scenario were limited to a maximum of five and there was no time constraint. Figure 3 summarizes the different configurations used.

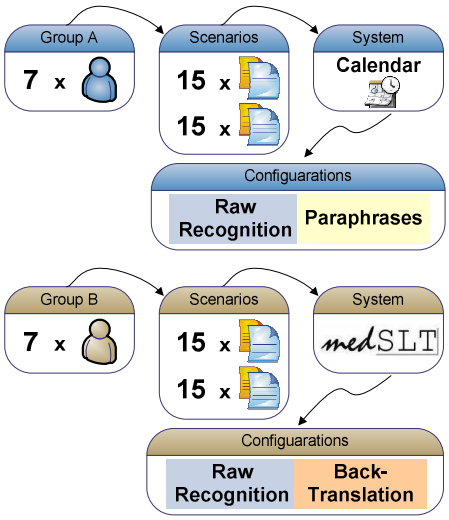

Figure 3: Configuration of the experiments

The only difference in MedSLT was that instead of using the paraphrases, the user was presented with the back-translation. Furthermore the scenarios follow a dialogue flow so that information can be accumulated from previous steps of the interaction. The users were native French speakers, with proficient computer skills and the scenarios were created as a translation task between English to French.

All subjects used a wired headset, which from our previous studies (Tsourakis et al, 2009, 2008) showed superior performance compared to a Bluetooth headset or to the onboard microphone.

Before using the system each user had to fill a demographic questionnaire and upon completion of each experiment an evaluation questionnaire according to ITU-T Rec. P.851 (2003). They also had five minutes of introduction on using the system and familiarizing themselves with the offered functionalities. All experiments took place in an office environment, while participants were seated. Table 2 shows sample scenarios for Calendar and MedSLT application.

Sample Scenarios for Calendar :
"You want to know if you have a meeting this Friday morning."
"You have a meeting with Susan, but you don't remember when it is."
"You can't remember who was at your last meeting."
"You want to check whether you have a meeting on January 5th."
"You have a meeting with Mike next month, but you don't remember the date."
Sample Scenarios for MedSLT :
"Does the pain start in the morning?"
"In the evening?"
"Do you have pain in the forehead?"
"tea + relieve + headaches"
"pain + cause + cough"

Table 2: Sample scenarios for Calendar & MedSLT application.

4. Results and Discussion

In this section we will present the results of the evaluation with real users concerning both MedSLT and Calendar applications. We will split the presentation into two subsections, one subsection for objective measures and one for the subjective evaluation. All results will be followed by a short discussion.

4.1 Objective evaluation

4.1.1. Mental workload

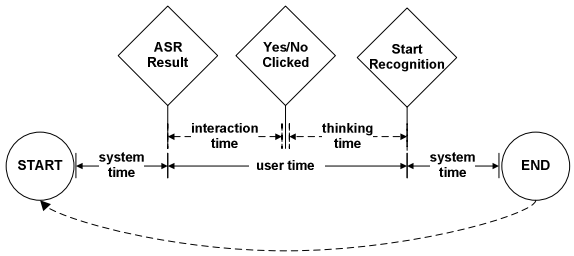

Figure 4: Decomposition of user/system time

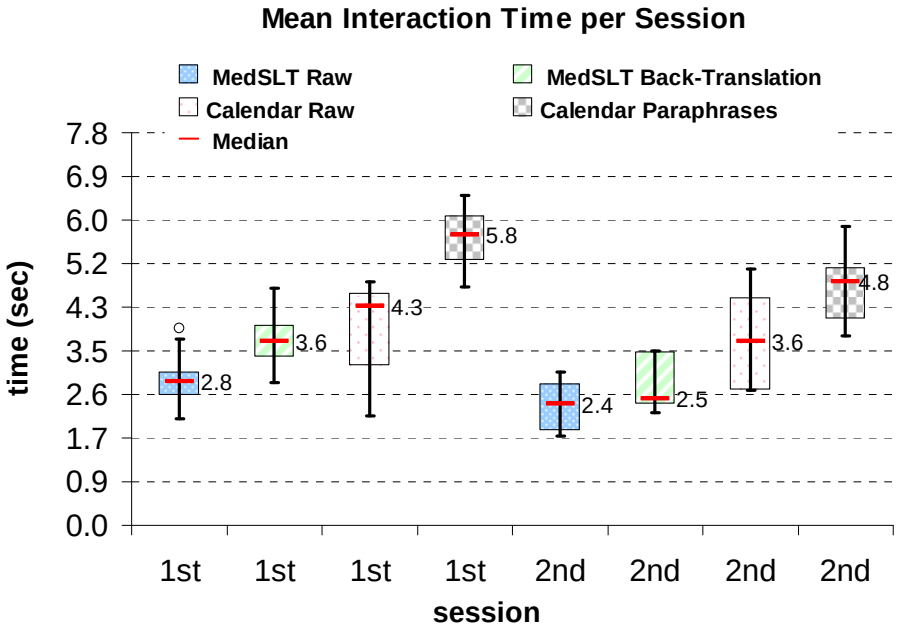

Figure 5: Mean interaction time per session

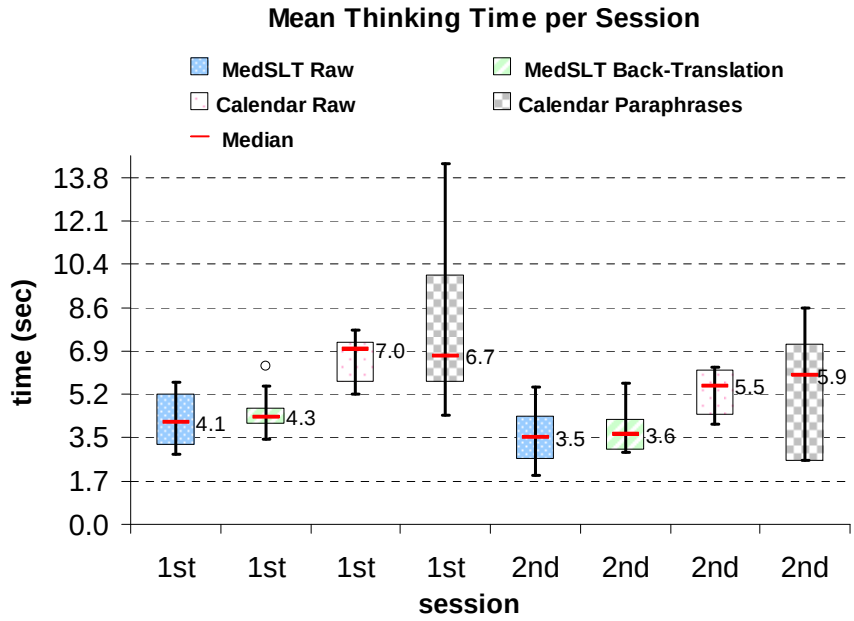

Figure 6: Mean thinking time per session

In order to quantify the mental effort we define "interaction time" as the time between the presentation of the result and its confirmation/rejection by the user. We define "thinking time" as the time spent by the user reading the scenario and formulating the corresponding question in his mind. We roughly consider this time as the interval between the confirmation/rejection of the previous scenario and the press of the recognition button. The decomposition of the time intervals is shown in Figure 4.

From the box-plots in Figure 5 we can observe something more or less expected. Users spend more time confirming their rephrased input. This has to do with the time spent reading a normally richer output and comprehending its semantic representation. We used one tail paired t-test to calculate statistical significance. On average it takes 1/3 more time (2.8 sec to 3.6 sec) ($t=5.58$, $df=6$, $p<0.0001$) in the first session for MedSLT and 1/2 (4.3 sec to 5.8 sec) ($t=7.06$, $df=6$, $p<0.0001$) for Calendar. In the second session we observe respectively a difference between 2.4 sec to 2.5 sec ($t=2.71$, $df=6$, $p<0.02$) and 3.6 sec to 4.8 sec ($t=4.69$, $df=6$, $p<0.002$). The users seem to get more familiar with the system as time passes. On the other hand they show a uniform behavior when they are presented with the raw recognition result.

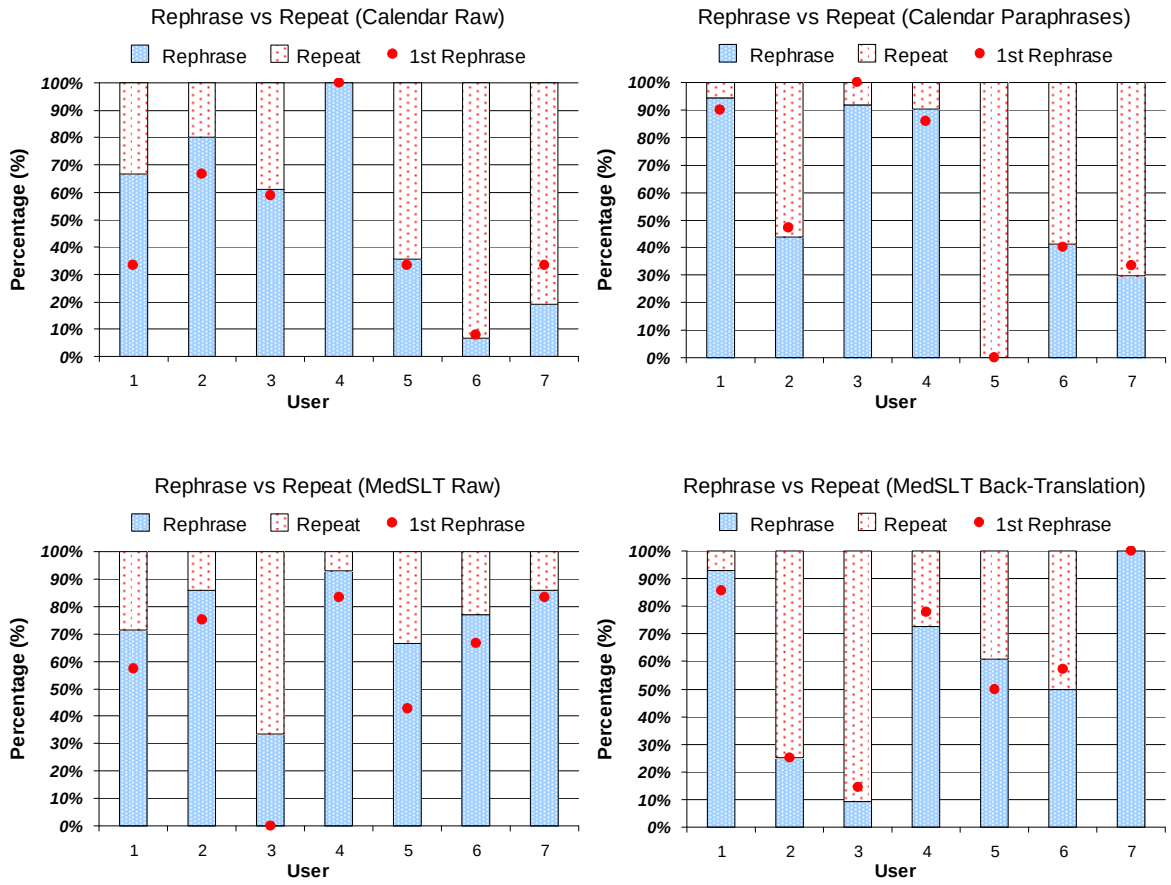

Figure 7: Rephrase vs Repeat preference in the four versions

The time needed for a user to pose a question to the system (thinking time), is presented in Figure 6. On average they spend almost the same time in each of the different pairs under examination. In the first session for example we observe a variation between 4.1 to 4.3 sec for MedSLT and 7 sec to 6.7 sec for Calendar. During the second session users become confident with the system and interact more rapidly. This also has to do with the descriptions of the scenarios, which are shorter. The uniformity of the values suggests that the rephrased output doesn't impose any additional workload for the users while formulating their input.

The statistical significance analysis reveals that we cannot reject the null hypothesis that mean values of thinking time are equal across the different pairs.

We should note that the average speaking rate for all users remained the same between the raw and the rephrased configurations. This suggests that rephrasing user's input doesn't affect the time spent by each subject uttering their questions.

4.1.2. Efficiency

In order to examine the utility of the paraphrases in user interaction we should check how easy it was to complete the task. The users had an average of 2.5 turns in the paraphrase configuration and 3 turns in the raw recognition configuration using the Calendar application. This was more or less expected as rephrasing user input helps hide unimportant recognition errors. We calculated that 12% of all interactions contained crucial recognition errors, which were eliminated in the rephrased output. Notice that this rate does not include minor recognition errors like substitutions between articles (e.g. "a" and "the").

As far as MedSLT is concerned, we observe a similar performance where 1.5 turns are needed in order to accept a result in both configurations. This high performance is probably due to the construction of the scenarios, which were simple translation tasks. This will be supported by the results of the ASR evaluation.

As a side product of this analysis we calculated the percentage, where users prefer to repeat a misrecognized sentence or rephrase it. The results are depicted in Figure 7 and correspond to the four versions under test. The subjects exhibit a strong preference in either strategy (rephrasing or repeating), as in most cases either of them prevails in the column. An interesting result is that this preference is extended to both versions that each user had to test. From 14 subjects only two of them (one in Group A, user No 2 and one in Group B, user No 2) decided to change strategy when introduced to the second version of the system. In order to accomplish a scenario, users may interact more than once. The dot in the plots represents the percentage of choosing to rephrase after the first misrecognition. It's an indication of what was their first preference after an error occurred and normally it's consistent with their prevailing preference.

4.1.3 Accuracy

Another issue that we tried to investigate is whether and to what extent users discard correct output or accept false output. We will only present results from the Calendar application as the high performance of MedSLT didn't offer the opportunity to conclude on concrete results.

As can be observed in Figure 8, despite the nature of the task (simple and clear) almost 1/10 (10%) of the interactions were erroneous. More errors came from false rejections rather than false acceptances. It is not clear however if these errors were caused by the user's negligence, due to the small display of the device, or due to the structure of the output (either paraphrases or raw recognition).

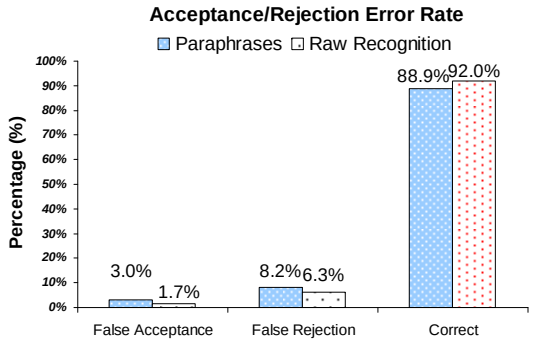

Figure 8: Output acceptance/rejection error rates for Calendar

4.1.4. Short-term learning

By the notion short-term we mean that the user learns from the examples and applies this information in the future. In order to quantify this process we examined the dialogue flow. We focus on examples where they could have used the rephrased output after a recognition error. We counted the number of interactions after an error where they used a useful pattern from the previous rephrased output. The number is low, equal to 5%, and includes only the use of some patterns and never full conversational sentences.

One indication of a learning process could also be offered by the out-of-vocabulary rates (OOV) presented in Table 3. In general we encountered low values, which suggest high grammar coverage. For Calendar the OOV is increased in the second session, whereas for MedSLT the opposite pattern applies. We believe that this happened due to ambiguity of the scenarios in the second session, which was stronger for Calendar.

Out-of-vocabulary			
Calendar Raw	**Calendar Paraphrases**	**MedSLT Raw**	**MedSLT Back-Trans.**
1st 1.52%	**1st** 1.95%	**1st** 2.15%	**1st** 2.18%
2nd 2.35%	**2nd** 2.77%	**2nd** 1.21%	**2nd** 1.6%

Table 3: Out-of-vocabulary rates for all versions during the first and the second session

In our experiments it's difficult to quantify the free recall as depicted in (Clark & Clark, 1977). Each scenario is a new task that formally doesn't demand the recall of any previous knowledge. We can deduce that users definitely learn what to say and become familiar with the system as time pass by. They exhibit though strong preference on the patterns that worked for them before and they seem reluctant to rephrase the proposed system output even if it's very close to the meaning of what they said. This is consistent with other studies that show that it is easier for people to model both the length and the vocabulary of a terse computer output than of a conversational one (Zoltan-Ford, 1991).

4.1.5. ASR performance

In order to quantify the speech recognition performance we use the following metrics: the Word Error Rate (WER) and the Sentence Error Rate (SER). SER is, as usual, defined as the proportion of utterances where at least one word is misrecognized. The calculations included both in and out of coverage data.

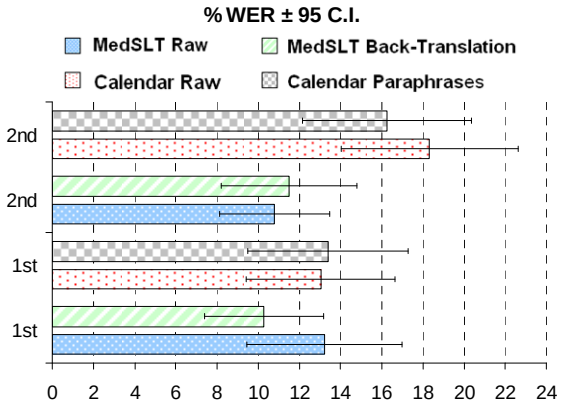

Figure 9: WER for the two configurations & systems during the first and the second session

In Figure 9 we present the plots of WER for each system and configuration, where we used a 95% confidence interval that was calculated after a per-utterance bootstrap resampling (Bisani & Ney, 2004). The results imply that statistically there is a uniform performance when users interact either with the raw or with the rephrased configuration. In other words there isn't enough evidence to suggest that rephrasing increase WER. Calendar version during the second session presents, as expected the highest values. The higher task completion rate for MedSLT compared to Calendar is consistent with the calculations of WER. The ambiguity introduced by the scenarios during the second round of the Calendar experiments seems to have an impact in performance. We shouldn't also forget that the subjects in Calendar were non native English speakers. Finally, the SER presented in Figure 10, denotes that for Calendar, 40% to 50% of the sentences had an error and this number drops to around 25% for MedSLT. This is perhaps a more convincing indication for the difference across systems and configurations for the task completion rate presented in section 4.1.2.

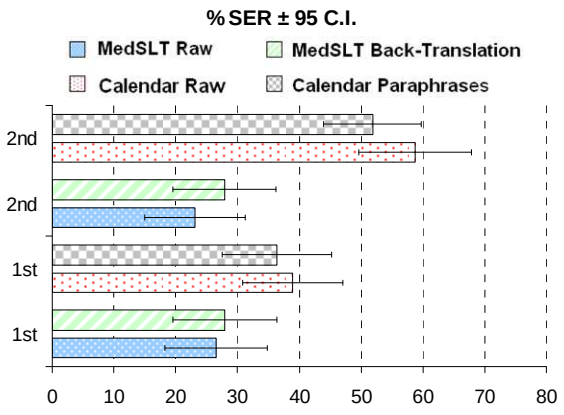

Figure 10: SER for the two configurations & systems during the first and the second session

A side product of this analysis revealed that for non native English speakers the choice of recognizing with the proper acoustic models is very important. The American acoustic models provided better results for some users whereas the British one were more appropriate for others. All experiment took place with the Nuance v8.5.0 recognition engine.

4.2 Subjective evaluation

The evaluation included an exit interview with a detailed questionnaire to measure the subjective opinion of the users for each system and configuration. Besides the written questionnaires we had a short discussion with all participants upon completion of each experiment.

Users didn't seem to favor one system over the other. In a five point Likert scale (Bad, Poor, Fair, Good, Excellent) they expressed their overall impression for both systems as "Good (4)" (Figure 11). For MedSLT there is a stronger tendency towards "Excellent (5)", as the interquartile range of the corresponding box-plots is above level 4.

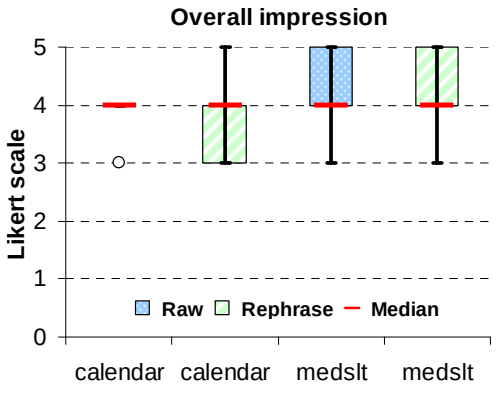

Figure 11: What was your overall impression of the system?

In Figures 12-13 we present two characteristics where users expressed different opinions. If we consider the Calendar application we observe the following. As a matter of comfort users seem to prefer the raw recognition, as most of the time they see an output very close if not identical to that which they originally gave.

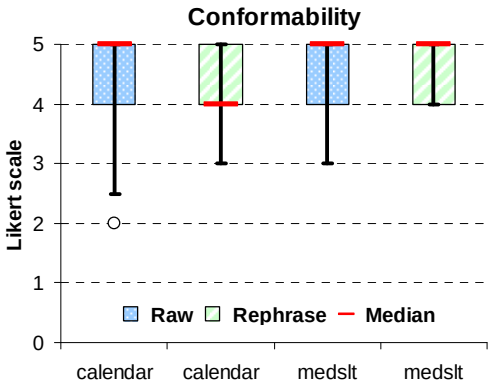

Figure 12: You were comfortable working with the system?

From our informal observations many of them felt initially awkward with the structure of their rephrased input. Calendar rephrase mechanism is different compared to MedSLT, by using abbreviations for months, introducing time intervals etc. The additional workload is already reported in the objective evaluation, as all versions outperform the Calendar-Paraphrases one. This had probably an impact to conformability.

One the other hand, as presented in Figure 13, they feel more confident in the system for dealing with misunderstandings when paraphrases are used. We suppose that the richer and detailed output gave them a second chance to reconsider their own input, which could be initially ambiguous. In any case the natural language processing might make them think that something clever takes place in the backend.

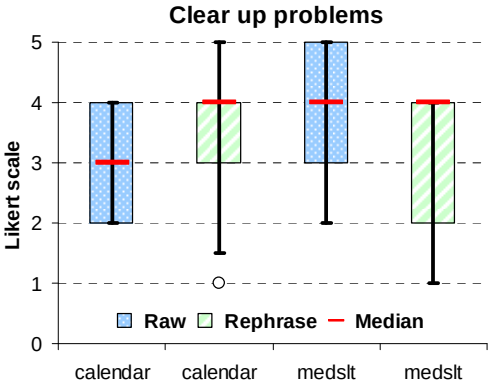

Figure 13: Misunderstandings could be cleared up easily?

Due to lack of establishing statistical significance for the subjective evaluation we won't extrapolate general conclusions. Conversely, we can rely on these results as empirical data and combine them with the discussions we had upon completion of each experiment. The assumptions reported in this work were also an outcome from the exit interviews.

Finally, we should take into account that the system didn't provide actual results. Whenever a user judged the raw recognition or the paraphrase as correct, he would also expect a correct query result. This is not the case however as hidden ambiguities in the raw recognition, revealed in the paraphrase or in back-translation, may offer false results. These rephrasing mechanisms seemed to the users as much a hindrance as an asset although their utility was not revealed in its full extent.

5 Conclusion

In this work we tried to investigate the effects of rephrasing as a confirmation mechanism of the user's input. We argue that for specific kinds of applications it is better to disambiguate the input before proceeding in the interaction. Especially in our case we can take advantage of the screen compared to a telephone based spoken language system. We utilized two applications running on a mobile platform that imposes different interaction to the users. This may be affected by the offered smaller display, the interaction with a stylus pen etc. We worked in an office environment albeit our study can be extended in different condition like outdoor testing or users on the move.

Our work was based on measurement of the additional workload, the efficiency and the performance of the system along with the satisfaction of the end users. We observed that as they become more familiar with the application, the time needed to process the output and re-interact with the system is reduced. In the worst case it's comparable with the time intervals presented for the raw recognition version. We can therefore state that rephrasing doesn't seem to impose additional work load concerning the time needed for a new interaction. The additional time for confirmation is strongly influenced by the structure of the rephrased output. The "soft" rephrasing mechanism of MedSLT compared to Calendar demands in general less time for confirmation.

We didn't notice more errors occurring when the rephrased output was used, although we would expect some kind of learning process. Perhaps in the context of this work there was not enough time for this process to occur considering also the conversational nature of the output, which does seem to be applicable. Probably another long-term evaluation could be germane for this kind of task.

From the overall impression of the users they seem to be equally satisfied with the raw recognition and the rephrased version. Some differences presented in the subjective analysis along with exit interviews and unofficial observation, make us conclude on issues concerning the conformability, where the raw version seems to be preferable. Conversely for the ability to clear up problems the rephrase version seems to be more applicable.

6 Acknowledgements

We would like to thank Nokia and its University Donation Program for offering two N810 Internet Tablets for use in the current work.

7 References

Bisani M. and Ney H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 409-411, Montreal, Canada.

Bohus D. and Rudnicky A. (2005). Constructing accurate beliefs in spoken dialog systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 272-277, San Juan, Puerto Rico.

Brennan S. and Ohaeri J.O. (1994). Effects of message style on user's attribution toward agents. *In Proceedings of CHI'94 Conference Companion Human Factors in Computing Systems*, pages 281–282. ACM Press.

Bouillon P., Rayner M., Chatzichrisafis N., Hockey B.A., Santaholma M., Starlander M., Nakao Y., Kanzaki K., and Isahara H. (2005). A generic multi-lingual open source platform for limited-domain medical speech translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, pages 50–58, Budapest, Hungary.

Clark H. H. and Clark E. V. (1977). Psychology and language: An introduction to psycholinguistics. New York: Harcourt Brace Jovanovich.

ITU-T Rec. P.851. 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.

Nass C. and Brave S. (2005). Should voice interfaces say "I"? Recorded and synthetic voice interfaces' claims to humanity. Chapter 10, pages 113–124. The MIT Press, Cambridge.

Rayner M., Hockey B.A., and Bouillon P. (2006). Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler. CSLI Press, Chicago.

Shin J., Narayanan S., Gerber L., Kazemzadeh A. and Byrd D. (2002). Analysis of user behavior under error conditions in spoken dialogs. *In Proceedings of ICSLP*, 2002, Denver, Colorado, USA.

Tsourakis N., Bouillon P., and Rayner M. (2009). Design issues for a bidirectional medical speech translator. In *Proceedings of SiMPE 2009*, Bonn, Germany.

Tsourakis N., Georgescul M., Bouillon P., and Rayner M. (2008). Building mobile spoken dialogue applications using Regulus. In *Proceedings of LREC 2008*, Marrakesh, Morocco.

Walker M.A., Litman D.J., Kamm C.A and Abella A. (1998). Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. In Computer Speech and Language.

Weegels M. (2004). User's Conceptions of Voice-Operated Information Services. *International Journal of Speech Technology, 3(2):75-82.*

Zoltan-Ford E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527–547.