CHAPTER 7

Higgins evaluation

In the previous chapter, the HIGGINS system was presented with a focus on concept-level error handling in the robust interpreter PICKERING and the discourse modeller GALATEA. In this chapter, two evaluations of these HIGGINS components are presented based on different data sets. The first is an evaluation of PICKERING, performed before the complete system was put together, and it is intended to explore how some of the robustness features implemented in PICKERING contribute to the performance. In the second evaluation, naive users were allowed to interact with the complete HIGGINS system. This evaluation is focussed on the performance of GALATEA, but also on users' behaviour when faced with fragmentary clarification requests.

7.1 PICKERING evaluation

A robust interpreter may be too allowing and, as a result, find incorrect interpretations. Hence, it cannot be taken for granted that the techniques PICKERING employs (as described in 6.4.2) increase the performance of the interpreter under error conditions. To check for this, PICKERING was evaluated with respect to how it performs under different error conditions, and how its performance depends on the robustness techniques. The evaluation included data with varying degrees of errors, to ensure that the interpreter works well under perfect conditions and degrades gracefully in the presence of errors.

7.1.1 Method and data

PICKERING was evaluated before the complete system was built. In order to collect data, eight subjects, all native speakers of Swedish, were given the task of moving around in the virtual 3D-city while they described their positions relative to objects in their surroundings. This

resulted in 340 utterances similar to those used during the positioning phases in the pedestrian navigation domain. The utterances were transcribed and a PICKERING grammar was written to cover the syntax and semantics that were deemed to be relevant to the domain and the objects and properties contained in the database.

For evaluation, 16 new subjects were recorded and their utterances transcribed using the same procedure. Due to time limitations, a test set of 100 of their utterances were manually annotated with semantics to create a gold standard for the evaluation. The average utterance length of the utterances was 11 words. The utterances were recognised using the KTH LVCSR speech recogniser (Seward, 2003) with a trigram language model. In order to study the effect of varying levels of errors in the ASR results on the PICKERING performance, the ASR was run repeated times on the material with the beam pruning level[9] set at different values. 18 pruning levels where used; the lowest pruning yielded an average WER rate of 34.2% and the highest 95.3%.

Next, the transcriptions and the ASR results were processed by PICKERING and compared to the gold standard in order to study the effects of ASR errors. The following interpreter parameters were systematically varied and combined:

- Agreement:
  - Weak: Agreement inside phrases (mostly congruence in nominal phrases) was preferred, but not required
  - Strong: Agreement inside phrases was required
- Permitted insertions inside phrases:
  - 0, 1, 2 or unlimited.

To simplify comparison of the results with the gold standard, an approximation was made by flattening the semantic trees to form lists of minimal concepts (i.e., nodes in the trees). On average, there were 32 such concepts per utterance. The lists were compared to the gold standard using standard WER calculation to get a concept error rate (CER). The percentage of concept insertions and deletions were calculated separately. Substitutions were counted as a combination of a deletion and an insertion, and were added to both groups.

The lexicon of the interpreter was also used to identify keywords in the utterances (i.e., words carrying semantics). This way, it was possible to estimate performance of PICKERING compared to that of a simple keyword spotter.

7.1.2 Results

For transcribed utterances (i.e., WER=0), the CER was 20%, with 12.6% deletions and 9.9% insertions (using the default setting, arbitrarily chosen when PICKERING was implemented: 2

[9] The beam pruning level determines how much of the ASR HMM space is explored for the optimal solution. A high pruning level means that less space is explored, which increases the speed of the ASR but reduces the accuracy.

allowed insertions, weak agreement). Figure 7.1 shows how the performance for insertions and deletions (Y-axis) varies as WER increases. The X-axis represents the mean WER for the different beam pruning settings. The figure also includes the ASR insertion/deletion performance for keywords.

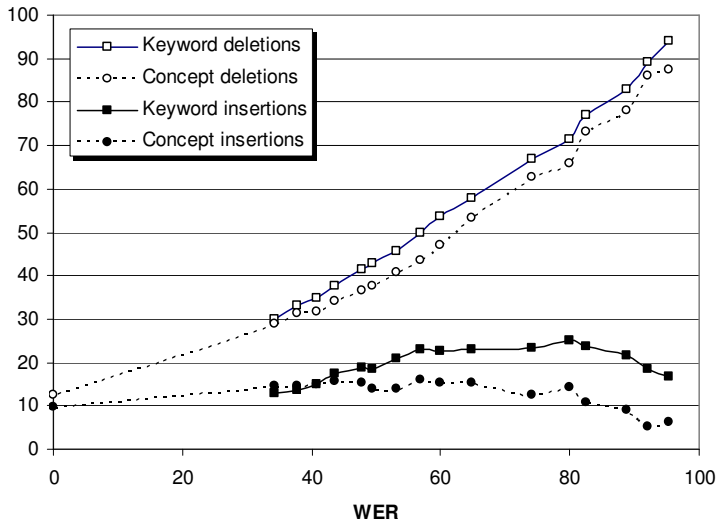

Figure 7.1: The percentage of insertions and deletions for keywords and concepts, depending on mean WER.

The figure shows that deletions rise steadily both for concepts and keywords, while insertions rise to a certain peak for keywords, but not for concepts. This is probably due to the fact that PICKERING may ignore erroneous content words in the input that do not fit into any syntax, as described in 6.4.2.1. To test the significance of the differences between keywords and concepts, the areas between the lowest and the highest WER's under the curves shown in the figure were compared for all utterances. There were significant differences between keywords and concepts for both insertions and deletions (two-tailed paired t-tests; dF=106; $p<0.05$).

To compare the effects of different parameter settings on the general robustness of PICKERING, a mean relative CER score (CER minus WER) was calculated for each parameter setting, which is shown in Figure 7.2. In order for this score to reflect the general performance over the whole WER span (34-95%), the average of all values of CER (relative to WER) along this span was calculated.

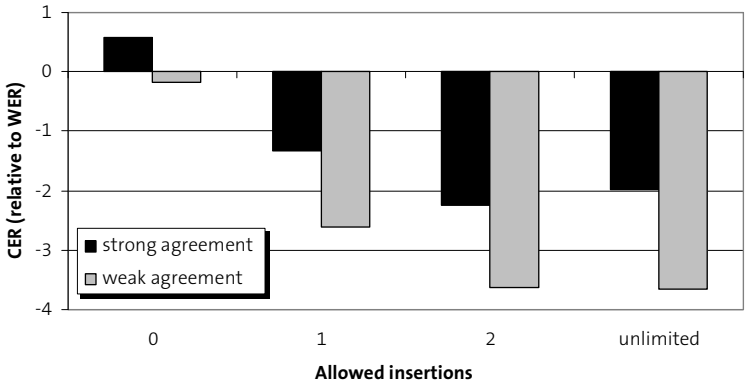

Figure 7.2: Mean CER (relative to WER) depending on parameter settings.

There were main effects for both strength of agreement and number of allowed insertions (two-way repeated measures ANOVA; dF=106; $p<0.05$). Post tests revealed that there were significant differences between 0, 1 and 2 insertions ($p<0.05$), but the performance did neither decrease nor increase when more than two insertions were allowed.

The results indicate that PICKERING generalises well when applied to unseen utterances, within the limited domain, produced by new speakers. As the WER of the ASR output increases, the set of robustness techniques utilised leads to graceful degradation of the interpretation results. The two techniques used to relax the CFG-constraints that were tested – allowing non-agreement and insertions – both improved performance. Since allowing insertions increases the size of the chart during parsing and thus slows down the computation, there seems to be no reason for allowing more than two insertions (given the circumstances).

7.2 GALATEA evaluation

To evaluate the performance of the HIGGINS system in general and GALATEA in particular, the complete system was tested with naive subjects using the system in a laboratory setting with given scenarios. The analysis of the results should be viewed as a proof-of-concept, to confirm that the system can interact with naive users and perform reasonably well. No comparative evaluation, involving other systems or different settings, is made. It is not possible to evaluate all aspects of the error handling techniques; the analysis of the results will focus on robustness, ellipsis resolution and fragmentary clarification.

7.2.1 Method

7.2.1.1 Subjects

16 subjects participated in the evaluation, all native speakers of Swedish. They were 7 women and 9 men, ranging in age from 24 to 63 years (38 on average). Four of the subjects had some experience of speech technology, but no experience of dialogue system design.

7.2.1.2 Procedure

Each subject was given the task of finding the way to a given goal in a virtual city, by talking to a computer to get route directions. They were told that they needed to tell the computer where they wanted to go, that the computer had no knowledge of their current position, but that the computer had a very detailed map of the city, and it was able to locate them if they described their surroundings. The subjects were not given any information about what kind of expressions the system could handle, or the level of detail of the system's knowledge. Four subsequent scenarios (goals) were given to each subject, resulting in a total of 64 dialogues.

The subject was placed in a sound-proofed room in front of a computer screen, where the 3D model of the virtual city was displayed from a first-person perspective, as seen in Figure 6.1. The subject spoke to the system through a headset and used a mouse to control the movement in the virtual city. During the interaction, the experiment conductor was sitting in another room, overlooking the interaction through a window, and did not answer any questions from the subject. For each scenario, there was a time limit of 10 minutes to reach the goal. After each scenario, the subjects filled out a questionnaire about their experience of the interaction. However, the results from this survey will not be used in the current analysis.

7.2.1.3 Data, ASR and TTS

The system was trained and configured mainly based on two different sets of data: the data presented in Chapter 4 and the data used for the evaluation of PICKERING. In addition to this, data were collected from four pilot sessions.

The collected data was used to write rules for PICKERING and GALATEA and to program the action manager, as well as to train the ASR language models. Since the ASR used for the PICKERING evaluation presented above did not support word confidence scores, an off-the-shelf ASR with such support was used. A trigram class-based language model was used, trained on the 1500 utterances that had been collected. Words that only occurred once in the training material were pruned, resulting in a vocabulary size of approximately 600 words.

As described in 6.7.7, a diphone Swedish male MBROLA voice was used for TTS and a very simple model for generating fragmentary clarification requests was used.

7.2.2 Results

7.2.2.1 Annotation

The dialogues were transcribed and annotated by one annotator. The segmentation of units for assigning features on the "utterance" level is not straightforward. One segmentation was based on the ASR endpoint detector – each ASR result was assigned a set of features. Another set of features was assigned to each CA, as segmented by the annotator. There was a pretty large discrepancy between these two methods for segmenting "utterances"; some CA's were not detected at all by the ASR, some were split over several ASR results, some ASR results contained several CA's. There were a total of 1894 ASR results and 2007 CA's, 1565 of the ASR results contained only one unsplit CA.

Both ASR results and CA's were annotated based on how well they were understood by the system. In the separate evaluation of PICKERING presented above, the commonly used measure of concept error rate (CER) was used. In this evaluation, however, the result that is to be assessed is an updated discourse model, including identified referents and enriched fragments. It would be too time-consuming to hand-craft a target discourse model to compare with for each utterance, at least if the whole data set is to be evaluated. To make the analysis more straightforward and the results easier to understand, five different levels of understanding were defined, similar to those used in Chapter 4. These levels are described in Table 7.1.

Table 7.1: The definitions of the understanding levels used in the annotation of ASR results and CA's.

Und.	Definition
FULLUND	All concepts, relevant to the domain and task, are fully understood by the system, including speech recognition, semantic interpretation, and discourse modelling. Note that the full propositional meaning must be understood; fragmentary utterances have to be correctly transformed into full propositions, and anaphoric expressions correctly resolved. This includes cases where the system asks for clarification on some of the concepts that it lacks confidence in.
PARTUND	Some (but not all) concepts are fully understood, according to the definition above. This includes cases where just a fragment of the utterance is interpreted, but the action manager uses this to pose a relevant question.
MISUND	Some (or all) concepts in the interpretation of the utterance are incorrect. This includes cases where the system asks for clarification on incorrect concepts.
NONUND	No concepts are understood. This includes cases where PICKERING might correctly interpret a fragment of the utterance (or a complete fragmentary utterance), but GALATEA fails to resolve the ellipsis, and the action manager fails to use the fragment to pose a relevant question.
NONHEAR	The CA is not detected at all. (Not applicable to ASR results.)

The CA's were also annotated based on the form and function of the utterance, but the scheme for this will not be described in detail here. An excerpt from the collected data with the different understanding levels for each CA is shown in Table 7.2.

Table 7.2: A dialogue excerpt from the evaluation data, translated from Swedish. The third column shows ASR results in which word confidence scores are represented with different shades of grey, where dark means high confidence. The fourth column shows the understanding level that was assigned to the CA by the annotator.

Turn	Transcription	ASR result	CA und.
U.1	I– I want to see a bus stop.	**YES I WANT TO GO TO A BUS STOP**	FULLUND
S.2	*Ok, where are you?*		
U.3	Eh, I stand in front of a a wooden house with three storeys.	**NOW IS I STAND IN FRONT OF A WOODEN HOUSE WITH THREE STOREYS**	FULLUND
S.4	*Three?*		
U.5	Yes	**YES**	FULLUND
S.6	*Ok, describe something more.*		
U.7	I see a sign that reads ...	**I SEE A SIGN THAT READS**	PARTUND (with U.7b)
S.8	*Ok ...*		
U.7b	Må – Måbär street.	[undetected]	(part of U.7)
S.9	*Please wait ...*		
S.10	*Can you see a brick building on your left?*		
U.12	No,	**NOW ON MY RIGHT**	NONUND
U.13	on my right.		PARTUND
S.14	*What do you see on your right?*		
U.15	A brick building with three storeys.	**A BRICK BUILDING WITH THREE STOREYS**	FULLUND
S.16	*Ok, a brick building with three storeys, hm ...*		
U.17	That is correct.	[undetected]	NONHEAR
S.18	*I think I know where you are. You should see an orange building. Position yourself between it and the three storey building made of wood.*		
U.19	Ok	**OK**	FULLUND

7.2.2.2 General results

Of the 64 tasks, 50 were completed within the time limit of 10 minutes. The tasks that succeeded took 4.3 minutes on average.

On average, there were 3.8 words per CA. This average figure is weighted down by the frequent use of short acknowledgements (like "yes" and "ok") during the route description phases of the dialogues, as well as the high number of fragmentary utterances. For assertions, the average number of words per CA were 7.3. Of all words spoken by the subjects, 1.6 % were out-of-vocabulary, not counting truncated words.

The ASR results had an average word error rate (WER) of 23.6%. Table 7.3 shows the distribution of understanding levels for ASR results and CA's, as well as for the ASR results containing just one CA, and for the subset of these that had a WER of 0%.

Table 7.3: The number of instances and distribution of understanding for ASR results and CA's. The fourth column shows ASR results which contain one unsplit CA. The fifth column shows the subset of these that had a WER of 0%. The understanding levels are defined in Table 7.1.

	ASR results	**CA's**	**ASR res.=CA**	**ASR res.=CA 0% WER**
Instances	1894	2007	1565	1033
FULLUND	65.4 %	66.2 %	73.8 %	92.9 %
PARTUND	9.2 %	5.8 %	5.0 %	1.6 %
MISUND	10.1 %	8.3 %	9.5 %	1.5 %
NONUND	15.3 %	11.8 %	11.7 %	3.9 %
NONHEAR		8.0 %		

The rightmost column in Table 7.3 reflects the performance of PICKERING and GALATEA, that is, how well the rules written to handle the training data generalised to new data, given that there are no ASR errors. Considering the limited "training data", 92.9% FULLUND should be regarded as a promising performance. This also confirms that the ASR is the major source of errors.

The third column (*CA's*) shows that 8.0% of the CA's were not detected by the system at all. This is partly explained by the fact that there were a lot of feedback utterances from the user (such as "mhm") after separate display utterances from the system, which did not have enough intensity to trigger the voice activity detection. Another explanation is that the ASR was not allowed to deliver any results while the system was talking. It was never shut-off, but if a speech endpoint was detected and a system utterance was playing, no result was delivered. This was done to prevent some turn-taking problems that cannot yet be handled. An example of this is U.7b in Table 7.2, where the utterance ends while S.9 is spoken.

7.2.2.3 Robustness and early error detection

To study the robustness of PICKERING and GALATEA against ASR errors, the ASR results were divided in WER intervals. The distribution of understanding for these spans is shown in Figure 7.3.

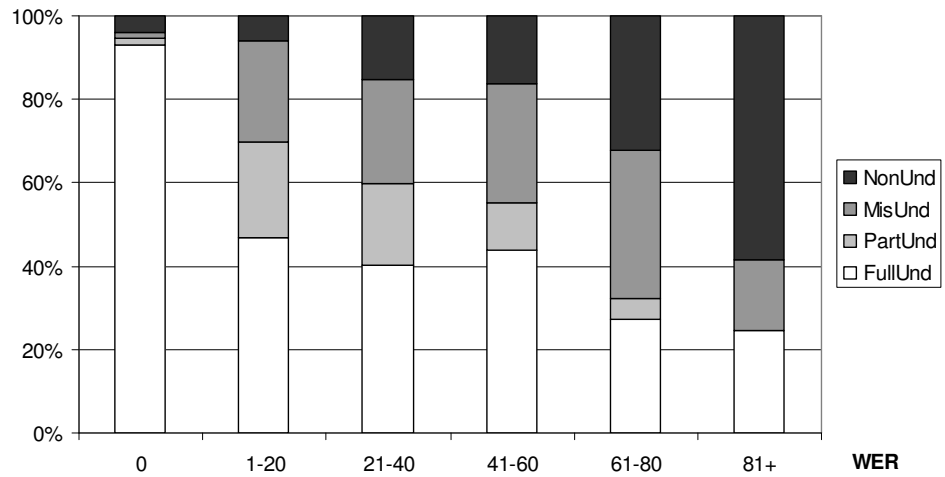

Figure 7.3: The distribution of understanding depending on the WER of the ASR result (rounded up).

The figure shows that the introduction of ASR errors immediately decreases the proportion of FULLUND by about 40%. Roughly half of this performance drop consists of some deleted concepts (PARTUND) and half by inserted concepts (MISUND). Interestingly, as the WER increases, the performance degrades gracefully up to a WER as high as 60%. Even with a WER above this, the proportion of misunderstandings seems to be stable, indicating an acceptable early error detection performance.

7.2.2.4 Ellipsis resolution

To find out how well GALATEA manages to resolve ellipsis, correctly recognised task-related complete CA's were grouped based on their form. Table 7.4 shows the distribution of understanding for some relevant forms. In this context, the form "fragment" includes adjectives, nouns, nominal phrases, propositional phrases, etc. As the table shows, fragmentary utterances – where ellipsis resolution is needed for full understanding – were almost as successful as assertions. For fragments, there was a larger proportion of partial understandings. These are utterances like U.13 in Table 7.2 which can be tricky to resolve correctly, but may be used by the system to ask a request like S.14, resulting in a partial understanding. Acknowledgements and yes/no-utterances also need ellipsis resolution. But, as the table shows, this is an easier task.

Table 7.4: The distribution of understanding for recognised task-related complete CA's of different forms, where WER=0%.

	Assertion	Fragment	"Ok"	"Yes"/"No"
FULLUND	90.1 %	88.6 %	100.0 %	98.2 %
PARTUND	1.6 %	8.4 %	0.0 %	0.0 %
MISUND	2.6 %	1.8 %	0.0 %	0.0 %
NONUND	5.8 %	1.2 %	0.0 %	1.8 %

7.2.2.5 Fragmentary clarification

There were a total of 94 fragmentary clarification requests in the data. Of these, 68.1% followed upon a correct hypothesis of the user's utterance and 31.9 % followed upon an incorrect hypothesis (i.e., a misunderstanding). The function and type of the user CA following the request were used to group the user reactions to the requests based on six different types. Table 7.5 shows the distribution of these types.

Table 7.5: Immediate user reaction to fragmentary clarification requests. The second column shows the distribution for cases where the request followed upon a correct hypothesis, and the third column cases where the request followed upon an incorrect hypothesis.

Reaction	Correct	Incorrect
"Yes"-answer.	59.4 %	0.0 %
"No"-answer.	3.1 %	40.0 %
A correction or elaboration, in the form of a fragment or assertion, as an answer to the request.	10.9 %	20.0 %
An utterance that relates to the request, but does not answer it.	4.7 %	6.7 %
A signal of non-understanding (such as "what? ").	9.4 %	6.7 %
The request is ignored.	12.5 %	26.7 %

Considering the large proportion of reactions that either ignore the request or signal non-understanding, fragmentary clarification requests seem to be hard for the users to understand. This may be explained partly by the fact that users may not expect such human-like behaviour from dialogue systems. Another explanation is that the clarification requests were sometimes used in contexts where a human would not have used them. A third explanation is that the prosodic model used to realise them (as described in 6.7.7) was very simplistic and not tested. This issue is explored in more depth in Chapter 9.

Fragmentary clarification requests seem to be even harder to understand after misunderstandings. This is of course due to the fact that such requests may not make sense to the user in some situations. For example, the request "red?" after a misrecognised "I want to go to a bus stop" may be perceived as inadequate.

There were many cases where the users ignored the clarification request after a correct recognition. However, it is possible that these should be interpreted as a "silent consent". Purver (2004) found that clarification requests in human-human dialogue are very often not answered (in 17-39% of the cases). Thus, the assumption taken here – that the concepts that are clarified must be confirmed to be considered as being correct – may be implausible.

Of the reactions that imply that the request was understood correctly by the user, it is interesting to note that far from all started with simple "yes" or "no" answers. Especially after misunderstandings, the user often corrects the system without starting with "no, ...". For many reactions, it is not obvious if they should be interpreted as answers to the preceding request, even if they relate to it. This supports the previously discussed assumption that clarification requests should not be treated as some sort of "subdialogue", but rather as a signal from the system that it lacks understanding, in the form of a request, which makes a fragmentary response possible to resolve. An interesting observation is the existence of some "no" answers after clarification requests based on correct understanding. These are cases where the user changes her mind, possibly because the system's request is interpreted as if it doubted the correctness of the user's description.

As discussed in Schlangen & Fernández (2007a), the frequent use of elaboration as a reaction to clarification requests may suggest that the users interpret them as concerning the understanding level and not the perception level (as discussed in 3.1.4), that is, they may have interpreted them as "do you really mean X", instead of "did you say X". If that was the case, it is possible that these interpretations were caused by the simplistic prosodic model used (which is explored further in Chapter 9).

Table 7.6 shows the distribution of understanding of the user reactions. For requests based on correct interpretations, the understanding of the responses seems to be similar to that of CA's in general (compare with Table 7.3). However, the performance is poorer for responses to requests based on misunderstandings, reflecting the fact that these were less predictable.

Table 7.6: The system's understanding of the user reactions to fragmentary clarification requests. The second column shows the distribution for cases where the clarified concepts were correctly understood by the system, and the third column cases where the clarified concepts were incorrect (i.e., after a misunderstanding).

Understanding	Correct	Incorrect
FULLUND	67.2 %	43.3 %
PARTUND	3.1 %	13.3 %
MISUND	6.3 %	6.7 %
NONUND	15.6 %	30 %
NONHEAR	7.8 %	6.7 %

7.2.2.6 Late error detection

The log files from the action manager showed that there were a total of 78 cases where there was not any place where the user could be when matching the discourse model against the database – indicating that a misunderstanding had occurred. In 45 of these cases, removing concepts with low grounding status made it possible for the system to continue positioning the user. This looks promising, but it does not tell us how many of the correct or incorrect concepts actually were removed. Future work will focus on answering this question, as well as finding methods for improving late error detection, as it is beyond the scope of this thesis.

7.3 Summary

An evaluation of the robust interpreter PICKERING indicates that it generalises well when applied to unseen utterances, within the limited domain, produced by new speakers. As the WER of the ASR output increases, the set of robustness techniques utilised leads to graceful degradation of the interpretation results. The two techniques used to relax the CFG-constraints that were tested – allowing non-agreement and insertions – both improved performance. Allowing an unlimited number of insertions into syntactical structures caused neither decline nor increase in accuracy.

An evaluation of the complete HIGGINS system showed that the performance of GALATEA and the rest of the system looks promising, not only when utterances are correctly recognised, but also when ASR errors are introduced. There have previously not been many studies on the use of fragmentary clarification requests in spoken dialogue systems interacting with real users. The results from this evaluation show that users may have difficulties understanding these requests, especially after incorrect speech recognition hypotheses, and that a more elaborate model of when to use them, how to realise them, and how to understand the user's reaction to them is needed.