Dependencies between Student State and Speech Recognition Problems in Spoken Tutoring Dialogues

Mihai Rotaru
University of Pittsburgh
Pittsburgh, USA
mrotaru@cs.pitt.edu

Diane J. Litman
University of Pittsburgh
Pittsburgh, USA
litman@cs.pitt.edu

Abstract

Speech recognition problems are a reality in current spoken dialogue systems. In order to better understand these phenomena, we study dependencies between speech recognition problems and several higher level dialogue factors that define our notion of student state: frustration/anger, certainty and correctness. We apply Chi Square ($\chi 2$) analysis to a corpus of speech-based computer tutoring dialogues to discover these dependencies both within and across turns. Significant dependencies are combined to produce interesting insights regarding speech recognition problems and to propose new strategies for handling these problems. We also find that tutoring, as a new domain for speech applications, exhibits interesting tradeoffs and new factors to consider for spoken dialogue design.

1 Introduction

Designing a spoken dialogue system involves many non-trivial decisions. One factor that the designer has to take into account is the presence of speech recognition problems (**SRP**). Previous work (Walker et al., 2000) has shown that the number of SRP is negatively correlated with overall user satisfaction. Given the negative impact of SRP, there has been a lot of work in trying to understand this phenomenon and its implications for building dialogue systems. Most of the previous work has focused on lower level details of SRP: identifying components responsible for SRP (acoustic model, language model, search algorithm (Chase, 1997)) or prosodic characterization of SRP (Hirschberg et al., 2004).

We extend previous work by analyzing the relationship between SRP and higher level dialogue factors. Recent work has shown that dialogue design can benefit from several higher level dialogue factors: dialogue acts (Frampton and Lemon, 2005; Walker et al., 2001), pragmatic plausibility (Gabsdil and Lemon, 2004). Also, it is widely believed that user emotions, as another example of higher level factor, interact with SRP but, currently, there is little hard evidence to support this intuition. We perform our analysis on three high level dialogue factors: frustration/anger, certainty and correctness. Frustration and anger have been observed as the most frequent emotional class in many dialogue systems (Ang et al., 2002) and are associated with a higher word error rate (Bulyko et al., 2005). For this reason, we use the presence of emotions like *frustration* and *anger* as our first dialogue factor.

Our other two factors are inspired by another contribution of our study: looking at *speech-based computer tutoring* dialogues instead of more commonly used information retrieval dialogues. Implementing spoken dialogue systems in a new domain has shown that many practices do not port well to the new domain (e.g. confirmation of long prompts (Kearns et al., 2002)). Tutoring, as a new domain for speech applications (Litman and Forbes-Riley, 2004; Pon-Barry et al., 2004), brings forward new factors that can be important for spoken dialogue design. Here we focus on *certainty* and *correctness*. Both factors have been shown to play an important role in the tutoring process (Forbes-Riley and Litman, 2005; Liscombe et al., 2005).

A common practice in previous work on emotion prediction (Ang et al., 2002; Litman and Forbes-Riley, 2004) is to transform an initial finer level emotion annotation (five or more labels) into a coarser level annotation (2-3 labels). We wanted to understand if this practice can im-

pact the dependencies we observe from the data. To test this, we combine our two emotion[1] factors (frustration/anger and certainty) into a binary emotional/non-emotional annotation.

To understand the relationship between SRP and our three factors, we take a three-step approach. In the first step, dependencies between SRP and our three factors are discovered using the Chi Square ($\chi 2$) test. Similar analyses on human-human dialogues have yielded interesting insights about human-human conversations (Forbes-Riley and Litman, 2005; Skantze, 2005). In the second step, significant dependencies are combined to produce interesting *insights* regarding SRP and to propose *strategies* for handling SRP. Validating these strategies is the purpose of the third step. In this paper, we focus on the first two steps; the third step is left as future work.

Our analysis produces several interesting insights and strategies which confirm the utility of the proposed approach. With respect to insights, we show that user emotions interact with SRP. We also find that incorrect/uncertain student turns have more SRP than expected. In addition, we find that the emotion annotation level affects the interactions we observe from the data, with finer-level emotions yielding more interactions and insights.

In terms of strategies, our data suggests that favoring misrecognitions over rejections (by lowering the rejection threshold) might be more beneficial for our tutoring task – at least in terms of reducing the number of emotional student turns. Also, as a general design practice in the spoken tutoring applications, we find an interesting tradeoff between the pedagogical value of asking difficult questions and the system's ability to recognize the student answer.

2 Corpus

The corpus analyzed in this paper consists of 95 experimentally obtained spoken tutoring dialogues between 20 students and our system **ITSPOKE** (Litman and Forbes-Riley, 2004), a speech-enabled version of the text-based WHY2 conceptual physics tutoring system (VanLehn et al., 2002). When interacting with ITSPOKE, students first type an essay answering a qualitative physics problem using a graphical user interface. ITSPOKE then engages the student in spoken dialogue (using speech-based input and output) to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. For recognition, we use the Sphinx2 speech recognizer with stochastic language models. Because speech recognition is imperfect, after the data was collected, each student utterance in our corpus was manually transcribed by a project staff member. An annotated excerpt from our corpus is shown in Figure 1 (punctuation added for clarity). The excerpts show both what the student said (the STD labels) and what ITSPOKE recognized (the ASR labels). The excerpt is also annotated with concepts that will be described next.

2.1 Speech Recognition Problems (SRP)

One form of SRP is the **Rejection**. Rejections occur when ITSPOKE is not confident enough in the recognition hypothesis and asks the student to repeat (Figure 1, $STD_{3,4}$). For our χ^2 analysis, we define the **REJ** variable with two values: **Rej** (a rejection occurred in the turn) and **noRej** (no rejection occurred in the turn). Not surprisingly, ITSPOKE also misrecognized some student turns. When ITSPOKE heard something different than what the student actually said but was confident in its hypothesis, we call this an **ASR Misrecognition** (a binary version of the commonly used Word Error Rate) (Figure 1, $STD_{1,2}$). Similarly, we define the **ASR MIS** variable with two values: **AsrMis** and **noAsrMis**.

Semantic accuracy is more relevant for dialogue evaluation, as it does not penalize for word errors that are unimportant to overall utterance interpretation. In the case of form-based information access spoken dialogue systems, computing semantic accuracy is straightforward (i.e. concept accuracy = percentage of correctly recognized concepts). In contrast, in the tutoring domain there are no clear forms with slots to be filled. We base our semantic accuracy on the "correctness" measure of the student turn. For each student turn, ITSPOKE interprets it and labels its correctness with regard to whether the student correctly answered the tutor question (see the labels between square brackets in Figure 1). We define **Semantic Misrecognition** as cases where ITSPOKE was confident in its recognition hypothesis and the correctness interpretation of the recognition hypothesis is different from the correctness interpretation of the manual transcript (Figure 1, STD_1). Similarly, we define the **SEM MIS** variable with two values: **SemMis** and **noSemMis**. The top part of Table 1 lists the distribution for our three SRP variables.

[1] We use the term "emotion" loosely to cover both affects and attitudes that can impact student learning.

3 Student State Annotation

Our definition of student state includes annotation on three dimensions (our three dialogue factors): Frustration/Anger/Hyperarticulation, Certainty and Correctness (see Table 1). The first two dimensions capture the student emotions and the last one the student correctness.

ITSPOKE: Good. Apply the same force to an object of mass M3, which is three times the mass of M1. What is the magnitude M3's acceleration in terms of A1 [acceleration of the object M1]?

STD_1: third of a one (***CERT=Uncertain, CRCT=C***) [*correct*]

ASR_1: they're it but a one (***AsrMis, SemMis***) [*incorrect*]

ITSPOKE: The acceleration of M3 [...] would be one-third the acceleration of M1. [...] In general for the same force what remains constant?

STD_2: weight (***CRCT=I***) [*incorrect*]

ASR_2: great (***AsrMis***) [*incorrect*]

ITSPOKE: Well... The product of the mass of a body and its acceleration remains constant for a given force. What is the relationship between a force acting on the box, the mass of the box, and its acceleration?

STD_3: newton's second law (***CERT=Uncertain, CRCT=PC***)

ASR_3: the cancel the fire (***Rej***)

ITSPOKE: Could you please repeat that?

STD_4: newton's second law (***FAH=FrAng, CRCT=PC***)

ASR_4: newton second long (***Rej***)

Figure 1: Human-Computer Dialogue Excerpt

The ***Frustration/Anger/Hyperarticulation*** dimension captures the perceived negative student emotional response to the interaction with the system. Three labels were used to annotate this dimension: frustration-anger, hyperarticulation and neutral. Similar to (Ang et al., 2002), because frustration and anger can be difficult to distinguish reliably, they were collapsed into a single label: frustration-anger (Figure 1, STD_4). Often, frustration and anger is prosodically marked and in many cases the prosody used is consistent with hyperarticulation (Ang et al., 2002). For this reason we included in this dimension the hyperarticulation label (even though hyperarticulation is not an emotion but a state). We used the hyperarticulation label for turns where no frustration or anger was perceived but nevertheless were hyperarticulated. For our interaction experiments we define the **FAH** variable with three values: **FrAng** (frustration-anger), **Hyp** (hyperarticulation) and **Neutral**.

The ***Certainty*** dimension captures the perceived student reaction to the questions asked by our computer tutor and her overall reaction to the tutoring domain (Liscombe et al., 2005). (Forbes-Riley and Litman, 2005) show that student certainty interacts with a human tutor's dialogue decision process (i.e. the choice of feedback). Four labels were used for this dimension: certain, uncertain (e.g. Figure 1, STD_1), mixed and neutral. In a small number of turns, both certainty and uncertainty were expressed and these turns were labeled as mixed (e.g. the student was certain about a concept, but uncertain about another concept needed to answer the tutor's question). For our interaction experiments we define the **CERT** variable with four values: **Certain**, **Uncertain**, **Mixed** and **Neutral**.

Variable	Values	Student turns (2334)
Speech recognition problems		
ASR	AsrMis	25.4%
MIS	noAsrMis	74.6%
SEM	SemMis	5.7%
MIS	noSemMis	94.3%
REJ	Rej	7.0%
	noRej	93.0%
Student state		
FAH	FrAng	9.9%
	Hyp	2.1%
	Neutral	88.0%
CERT	Certain	41.3%
	Uncertain	19.1%
	Mixed	2.4%
	Neutral	37.3%
CRCT	C	63.3%
	I	23.3%
	PC	6.2%
	UA	7.1%
EnE	Emotional	64.8%
	Neutral	35.2%

Table 1: Variable distributions in our corpus.

To test the impact of the emotion annotation level, we define the Emotional/Non-Emotional annotation based on our two emotional dimensions: neutral turns on both the FAH and the CERT dimension are labeled as neutral[2]; all other turns were labeled as emotional. Consequently, we define the **EnE** variable with two values: **Emotional** and **Neutral**.

Correctness is also an important factor of the student state. In addition to the correctness labels assigned by ITSPOKE (recall the definition of SEM MIS), each student turn was manually annotated by a project staff member in terms of their physics-related correctness. Our annotator used the human transcripts and his physics knowledge to label each student turn for various

[2] To be consistent with our previous work, we label hyperarticulated turns as emotional even though hyperarticulation is not an emotion.

degrees of correctness: correct, partially correct, incorrect and unable to answer. Our system can ask the student to provide multiple pieces of information in her answer (e.g. the question "Try to name the forces acting on the packet. Please, specify their directions." asks for both the names of the forces and their direction). If the student answer is correct and contains all pieces of information, it was labeled as correct (e.g. "gravity, down"). The partially correct label was used for turns where part of the answer was correct but the rest was either incorrect (e.g. "gravity, *up*") or omitted some information from the ideal correct answer (e.g. "gravity"). Turns that were completely incorrect (e.g. "no forces") were labeled as incorrect. Turns where the students did not answer the computer tutor's question were labeled as "unable to answer". In these turns the student used either variants of "I don't know" or simply did not say anything. For our interaction experiments we defined the **CRCT** variable with four values: **C** (correct), **I** (incorrect), **PC** (partially correct) and **UA** (unable to answer).

Please note that our definition of student state is from the tutor's perspective. As we mentioned before, our emotion annotation is for perceived emotions. Similarly, the notion of correctness is from the tutor's perspective. For example, the student might think she is correct but, in reality, her answer is incorrect. This correctness should be contrasted with the correctness used to define SEM MIS. The SEM MIS correctness uses ITSPOKE's language understanding module applied to recognition hypothesis or the manual transcript, while the student state's correctness uses our annotator's language understanding.

All our student state annotations are at the turn level and were performed manually by the same annotator. While an inter-annotator agreement study is the best way to test the ***reliability*** of our two emotional annotations (FAH and CERT), our experience with annotating student emotions (Litman and Forbes-Riley, 2004) has shown that this type of annotation can be performed reliably. Given the general importance of the student's uncertainty for tutoring, a second annotator has been commissioned to annotate our corpus for the presence or absence of uncertainty. This annotation can be directly compared with a binary version of CERT: Uncertain+Mixed versus Certain+Neutral. The comparison yields an agreement of 90% with a Kappa of 0.68. Moreover, if we rerun our study on the second annotation, we find similar dependencies. We are currently planning to perform a second annotation of the FAH dimension to validate its reliability.

We believe that our correctness annotation (CRCT) is reliable due to the simplicity of the task: the annotator uses his language understanding to match the human transcript to a list of correct/incorrect answers. When we compared this annotation with the correctness assigned by ITSPOKE on the human transcript, we found an agreement of 90% with a Kappa of 0.79.

4 Identifying dependencies using χ^2

To discover the dependencies between our variables, we apply the χ^2 test. We illustrate our analysis method on the interaction between certainty (CERT) and rejection (REJ). The χ^2 value assesses whether the differences between observed and expected counts are large enough to conclude a statistically significant dependency between the two variables (Table 2, last column). For Table 2, which has 3 degrees of freedom ((4-1)*(2-1)), the critical χ^2 value at a $p<0.05$ is 7.81. We thus conclude that there is a statistically significant dependency between the student certainty in a turn and the rejection of that turn.

Combination		Obs.	Exp.	χ^2
CERT – REJ				11.45
Certain – Rej	-	49	67	9.13
Uncertain – Rej	+	43	31	6.15

Table 2: CERT – REJ interaction.

If any of the two variables involved in a significant dependency has more than 2 possible values, we can look more deeply into this overall interaction by investigating how particular values interact with each other. To do that, we compute a binary variable for each variable's value in part and study dependencies between these variables. For example, for the value 'Certain' of variable CERT we create a binary variable with two values: 'Certain' and 'Anything Else' (in this case Uncertain, Mixed and Neutral). By studying the dependency between binary variables we can understand how the interaction works.

Table 2 reports in rows 3 and 4 all significant interactions between the values of variables CERT and REJ. Each row shows: 1) the value for each original variable, 2) the sign of the dependency, 3) the observed counts, 4) the expected counts and 5) the χ^2 value. For example, in our data there are 49 rejected turns in which the student was certain. This value is smaller than the expected counts (67); the dependency between Certain and Rej is significant with a χ^2 value of 9.13. A comparison of the observed counts and expected counts reveals the direction

(sign) of the dependency. In our case we see that certain turns are rejected *less* than expected (row 3), while uncertain turns are rejected *more* than expected (row 4). On the other hand, there is no interaction between neutral turns and rejections or between mixed turns and rejections. Thus, the CERT – REJ interaction is explained only by the interaction between Certain and Rej and the interaction between Uncertain and Rej.

5 Results - dependencies

In this section we present all significant dependencies between SRP and student state both within and across turns. Within turn interactions analyze the contribution of the student state to the recognition of the turn. They were motivated by the widely believed intuition that emotion interacts with SRP. Across turn interactions look at the contribution of previous SRP to the current student state. Our previous work (Rotaru and Litman, 2005) had shown that certain SRP will correlate with emotional responses from the user. We also study the impact of the emotion annotation level (EnE versus FAH/CERT) on the interactions we observe. The implications of these dependencies will be discussed in Section 6.

5.1 Within turn interactions

For the FAH dimension, we find only one significant interaction: the interaction between the FAH student state and the rejection of the current turn (Table 3). By studying values' interactions, we find that turns where the student is frustrated or angry are rejected more than expected (34 instead of 16; Figure 1, STD_4 is one of them). Similarly, turns where the student response is hyperarticulated are also rejected more than expected (similar to observations in (Soltau and Waibel, 2000)). In contrast, neutral turns in the FAH dimension are rejected less than expected. Surprisingly, FrAng does not interact with AsrMis as observed in (Bulyko et al., 2005) but they use the full word error rate measure instead of the binary version used in this paper.

Combination		Obs.	Exp.	χ^2
FAH – REJ				77.92
FrAng – Rej	+	34	16	23.61
Hyp – Rej	+	16	3	50.76
Neutral – Rej	-	113	143	57.90

Table 3: FAH – REJ interaction.

Next we investigate how our second emotion annotation, CERT, interacts with SRP. All significant dependencies are reported in Tables 2 and 4. In contrast with the FAH dimension, here we see that the interaction direction depends on the valence. We find that 'Certain' turns have less SRP than expected (in terms of AsrMis and Rej). In contrast, 'Uncertain' turns have more SRP both in terms of AsrMis and Rej. 'Mixed' turns interact only with AsrMis, allowing us to conclude that the presence of uncertainty in the student turn (partial or overall) will result in ASR problems more than expected. Interestingly, on this dimension, neutral turns do not interact with any of our three SRP.

Combination		Obs.	Exp.	χ^2
CERT – ASRMIS				38.41
Certain – AsrMis	-	204	244	15.32
Uncertain – AsrMis	+	138	112	9.46
Mixed – AsrMis	+	29	13	22.27

Table 4: CERT – ASRMIS interaction.

Finally, we look at interactions between student correctness and SRP. Here we find significant dependencies with all types of SRP (see Table 5). In general, correct student turns have fewer SRP while incorrect, partially correct or UA turns have more SRP than expected. Partially correct turns have more AsrMis and SemMis problems than expected, but are rejected less than expected. Interestingly, UA turns interact only with rejections: these turns are rejected more than expected. An analysis of our corpus reveals that in most rejected UA turns the student does not say anything; in these cases, the system's recognition module thought the student said something but the system correctly rejects the recognition hypothesis.

Combination		Obs.	Exp.	χ^2
CRCT – ASRMIS				65.17
C – AsrMis	-	295	374	62.03
I – AsrMis	+	198	137	45.95
PC – AsrMis	+	50	37	5.9
CRCT – SEMMIS				20.44
C – SemMis	+	100	84	7.83
I – SemMis	-	14	31	13.09
PC – SemMis	+	15	8	5.62
CRCT – REJ				99.48
C – Rej	-	53	102	70.14
I – Rej	+	84	37	79.61
PC – Rej	-	4	10	4.39
UA – Rej	+	21	11	9.19

Table 5: Interactions between Correctness and SRP.

The only exception to the rule is SEM MIS. We believe that SEM MIS behavior is explained by the "catch-all" implementation in our system. In ITSPOKE, for each tutor question there is a list of anticipated answers. All other answers are

treated as incorrect. Thus, it is less likely that a recognition problem in an incorrect turn will affect the correctness interpretation (e.g. Figure 1, STD_2: very unlikely to misrecognize the incorrect "weight" with the anticipated "the product of mass and acceleration"). In contrast, in correct turns recognition problems are more likely to screw up the correctness interpretation (e.g. misrecognizing "gravity down" as "gravity sound").

5.2 Across turn interactions

Next we look at the contribution of previous SRP – variable name or value followed by $_{(-1)}$ – to the current student state. Please note that there are two factors involved here: the presence of the SRP and the SRP handling strategy. In ITSPOKE, whenever a student turn is rejected, unless this is the third rejection in a row, the student is asked to repeat using variations of "Could you please repeat that?". In all other cases, ITSPOKE makes use of the available information ignoring any potential ASR errors.

Combination		Obs.	Exp.	χ^2
$ASRMIS_{(-1)}$ – FAH				7.64
$AsrMis_{(-1)}$ – FrAng	$-^t$	46	58	3.73
$AsrMis_{(-1)}$ – Hyp	$-^t$	7	12	3.52
$AsrMis_{(-1)}$ – Neutral	+	527	509	6.82
$REJ_{(-1)}$ – FAH				409.31
$Rej_{(-1)}$ – FrAng	+	36	16	28.95
$Rej_{(-1)}$ – Hyp	+	38	3	369.03
$Rej_{(-1)}$ – Neutral	-	88	142	182.9
$REJ_{(-1)}$ – CRCT				57.68
$Rej_{(-1)}$ – C	-	68	101	31.94
$Rej_{(-1)}$ – I	+	74	37	49.71
$Rej_{(-1)}$ – PC	-	3	10	6.25

Table 6: Interactions across turns (^{t} – trend, $p<0.1$).

Here we find only 3 interactions (Table 6). We find that after a non-harmful SRP (AsrMis) the student is less frustrated and hyperarticulated than expected. This result is not surprising since an AsrMis does not have any effect on the normal dialogue flow.

In contrast, after rejections we observe several negative events. We find a highly significant interaction between a previous rejection and the student FAH state, with student being more frustrated and more hyperarticulated than expected (e.g. Figure 1, STD_4). Not only does the system elicit an emotional reaction from the student after a rejection, but her subsequent response to the repetition request suffers in terms of the correctness. We find that after rejections student answers are correct or partially correct less than expected and incorrect more than expected. The $REJ_{(-1)}$ – CRCT interaction might be explained by the CRCT – REJ interaction (Table 5) if, in general, after a rejection the student repeats her previous turn. An annotation of responses to rejections as in (Swerts et al., 2000) (repeat, rephrase etc.) should provide additional insights.

We were surprised to see that a previous SemMis (more harmful than an AsrMis but less disruptive than a Rej) does not interact with the student state; also the student certainty does not interact with previous SRP.

5.3 Emotion annotation level

We also study the impact of the emotion annotation level on the interactions we can observe from our corpus. In this section, we look at interactions between SRP and our coarse-level emotion annotation (EnE) both within and across turns. Our results are similar with the results of our previous work (Rotaru and Litman, 2005) on a smaller corpus and a similar annotation scheme. We find again only one significant interaction: rejections are followed by more emotional turns than expected (Table 7). The strength of the interaction is smaller than in previous work, though the results can not be compared directly. No other dependencies are present.

Combination		Obs.	Exp.	χ^2
$REJ_{(-1)}$ – EnE				6.19
$Rej_{(-1)}$ – Emotional	+	119	104	6.19

Table 7: $REJ_{(-1)}$ – EnE interaction.

We believe that the $REJ_{(-1)}$ – EnE interaction is explained mainly by the FAH dimension. Not only is there no interaction between $REJ_{(-1)}$ and CERT, but the inclusion of the CERT dimension in the EnE annotation decreases the strength of the interaction between REJ and FAH (the χ^2 value decreases from 409.31 for FAH to a mere 6.19 for EnE). Collapsing emotional classes also prevents us from seeing any within turn interactions. These observations suggest that what is being counted as an emotion for a binary emotion annotation is critical its success. In our case, if we look at affect (FAH) or attitude (CERT) in isolation we find many interactions; in contrast, combining them offers little insight.

6 Results – insights & strategies

Our results put a spotlight on several interesting observations which we discuss below.

Emotions interact with SRP

The dependencies between FAH/CERT and various SRP (Tables 2-4) provide evidence that user's emotions interact with the system's ability

to recognize the current turn. This is a widely believed intuition with little empirical support so far. Thus, our notion of student state can be a useful higher level information source for SRP predictors. Similar to (Hirschberg et al., 2004), we believe that peculiarities in the acoustic/prosodic profile of specific student states are responsible for their SRP. Indeed, previous work has shown that the acoustic/prosodic information plays an important role in characterizing and predicting both FAH (Ang et al., 2002; Soltau and Waibel, 2000) and CERT (Liscombe et al., 2005; Swerts and Krahmer, 2005).

The impact of the emotion annotation level

A comparison of the interactions yielded by various levels of emotion annotation shows the importance of the annotation level. When using a coarser level annotation (EnE) we find only one interaction. By using a finer level annotation, not only we can understand this interaction better but we also discover new interactions (five interactions with FAH and CERT). Moreover, various state annotations interact differently with SRP. For example, non-neutral turns in the FAH dimension (FrAng and Hyp) will be always rejected more than expected (Table 3); in contrast, interactions between non-neutral turns in the CERT dimension and rejections depend on the valence ('certain' turns will be rejected less than expected while 'uncertain' will be rejected more than expected; recall Table 2). We also see that the neutral turns interact with SRP depending on the dimension that defines them: FAH neutral turns interact with SRP (Table 3) while CERT neutral turns do not (Tables 2 and 4).

This insight suggests an interesting tradeoff between the practicality of collapsing emotional classes (Ang et al., 2002; Litman and Forbes-Riley, 2004) and the ability to observe meaningful interactions via finer level annotations.

Rejections: impact and a handling strategy

Our results indicate that rejections and ITSPOKE's current rejection-handling strategy are problematic. We find that rejections are followed by more emotional turns (Table 7). A similar effect was observed in our previous work (Rotaru and Litman, 2005). The fact that it generalizes across annotation scheme and corpus, emphasizes its importance. When a finer level annotation is used, we find that rejections are followed more than expected by a frustrated, angry and hyperarticulated user (Table 6). Moreover, these subsequent turns can result in additional rejections (Table 3). Asking to repeat after a rejection does not also help in terms of correctness: the subsequent student answer is actually incorrect more than expected (Table 6).

These interactions suggest an interesting strategy for our tutoring task: favoring misrecognitions over rejections (by lowering the rejection threshold). First, since rejected turns are more than expected incorrect (Table 5), the actual recognized hypothesis for such turns turn is very likely to be interpreted as incorrect. Thus, accepting a rejected turn instead of rejecting it will have the same outcome in terms of correctness: an incorrect answer. In this way, instead of attempting to acquire the actual student answer by asking to repeat, the system can skip these extra turn(s) and use the current hypothesis. Second, the other two SRP are less taxing in terms of eliciting FAH emotions (recall Table 6; note that a SemMis might activate an unwarranted and lengthy knowledge remediation subdialogue). This suggests that continuing the conversation will be more beneficial even if the system misunderstood the student. A similar behavior was observed in human-human conversations through a noisy speech channel (Skantze, 2005).

Correctness/certainty–SRP interactions

We also find an interesting interaction between correctness/certainty and system's ability to recognize that turn. In general correct/certain turns have less SRP while incorrect/uncertain turns have more SRP than expected. This observation suggests that the computer tutor should ask the right question (in terms of its difficulty) at the right time. Intuitively, asking a more complicated question when the student is not prepared to answer it will increase the likelihood of an incorrect or uncertain answer. But our observations show that the computer tutor has more trouble recognizing correctly these types of answers. This suggests an interesting tradeoff between the tutor's question difficulty and the system's ability to recognize the student answer. This tradeoff is similar in spirit to the initiative-SRP tradeoff that is well known when designing information-seeking systems (e.g. system initiative is often used instead of a more natural mixed initiative strategy, in order to minimize SRP).

7 Conclusions

In this paper we analyze the interactions between SRP and three higher level dialogue factors that define our notion of student state: frustration/anger/hyperarticulation, certainty and correctness. Our analysis produces several interesting insights and strategies which confirm the

utility of the proposed approach. We show that user emotions interact with SRP and that the emotion annotation level affects the interactions we observe from the data, with finer-level emotions yielding more interactions and insights.

We also find that tutoring, as a new domain for speech applications, brings forward new important factors for spoken dialogue design: certainty and correctness. Both factors interact with SRP and these interactions highlight an interesting design practice in the spoken tutoring applications: the tradeoff between the pedagogical value of asking difficult questions and the system's ability to recognize the student answer (at least in our system). The particularities of the tutoring domain also suggest favoring misrecognitions over rejections to reduce the negative impact of asking to repeat after rejections.

In our future work, we plan to move to the third step of our approach: testing the strategies suggested by our results. For example, we will implement a new version of ITSPOKE that never rejects the student turn. Next, the current version and the new version will be compared with respect to users' emotional response. Similarly, to test the tradeoff hypothesis, we will implement a version of ITSPOKE that asks difficult questions first and then falls back to simpler questions. A comparison of the two versions in terms of the number of SRP can be used for validation.

While our results might be dependent on the tutoring system used in this experiment, we believe that our findings can be of interest to practitioners building similar voice-based applications. Moreover, our approach can be applied easily to studying other systems.

Acknowledgements

This work is supported by NSF Grant No. 0328431. We thank Dan Bohus, Kate Forbes-Riley, Joel Tetreault and our anonymous reviewers for their helpful comments.

References

J. Ang, R. Dhillon, A. Krupski, A. Shriberg and A. Stolcke. 2002. *Prosody-based automatic detection of annoyance and frustration in human-computer dialog.* In Proc. of ICSLP.

I. Bulyko, K. Kirchhoff, M. Ostendorf and J. Goldberg. 2005. Error-correction detection and response generation in a spoken dialogue system. *Speech Communication, 45*(3).

L. Chase. 1997. *Blame Assignment for Errors Made by Large Vocabulary Speech Recognizers.* In Proc. of Eurospeech.

K. Forbes-Riley and D. J. Litman. 2005. *Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus.* In Proc. of SIGdial.

M. Frampton and O. Lemon. 2005. *Reinforcement Learning of Dialogue Strategies using the User's Last Dialogue Act.* In Proc. of IJCAI Workshop on Know.&Reasoning in Practical Dialogue Systems.

M. Gabsdil and O. Lemon. 2004. *Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems.* In Proc. of ACL.

J. Hirschberg, D. Litman and M. Swerts. 2004. Prosodic and Other Cues to Speech Recognition Failures. *Speech Communication, 43*(1-2).

M. Kearns, C. Isbell, S. Singh, D. Litman and J. Howe. 2002. *CobotDS: A Spoken Dialogue System for Chat.* In Proc. of National Conference on Artificial Intelligence (AAAI).

J. Liscombe, J. Hirschberg and J. J. Venditti. 2005. *Detecting Certainness in Spoken Tutorial Dialogues.* In Proc. of Interspeech.

D. Litman and K. Forbes-Riley. 2004. *Annotating Student Emotional States in Spoken Tutoring Dialogues.* In Proc. of SIGdial Workshop on Discourse and Dialogue (SIGdial).

H. Pon-Barry, B. Clark, E. O. Bratt, K. Schultz and S. Peters. 2004. *Evaluating the effectiveness of Scot:a spoken conversational tutor.* In Proc. of ITS Workshop on Dialogue-based Intellig. Tutoring Systems.

M. Rotaru and D. Litman. 2005. *Interactions between Speech Recognition Problems and User Emotions.* In Proc. of Eurospeech.

G. Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication, 45*(3).

H. Soltau and A. Waibel. 2000. *Specialized acoustic models for hyperarticulated speech.* In Proc. of ICASSP.

M. Swerts and E. Krahmer. 2005. Audiovisual Prosody and Feeling of Knowing. *Journal of Memory and Language, 53*.

M. Swerts, D. Litman and J. Hirschberg. 2000. *Corrections in Spoken Dialogue Systems.* In Proc. of ICSLP.

K. VanLehn, P. W. Jordan, C. P. Rosé, et al. 2002. *The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing.* In Proc. of Intelligent Tutoring Systems (ITS).

M. Walker, D. Litman, C. Kamm and A. Abella. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*.

M. Walker, R. Passonneau and J. Boland. 2001. *Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems.* In Proc. of ACL.