Labeling Corrections and Aware Sites in Spoken Dialogue Systems

Julia Hirschberg† and Marc Swerts‡ and Diane Litman†

† AT&T Labs–Research
Florham Park, NJ, 07932 USA
{julia/diane}@research.att.com

‡ IPO, Eindhoven, The Netherlands,
and CNTS, Antwerp, Belgium
m.g.j.swerts@tue.nl

Abstract

This paper deals with user corrections and aware sites of system errors in the TOOT spoken dialogue system. We first describe our corpus, and give details on our procedure to label corrections and aware sites. Then, we show that corrections and aware sites exhibit some prosodic and other properties which set them apart from 'normal' utterances. It appears that some correction types, such as simple repeats, are more likely to be correctly recognized than other types, such as paraphrases. We also present evidence that system dialogue strategy affects users' choice of correction type, suggesting that strategy-specific methods of detecting or coaching users on corrections may be useful. Aware sites tend to be shorter than other utterances, and are also more difficult to recognize correctly for the ASR system.

1 Introduction

Compared to many other systems, spoken dialogue systems (SDS) tend to have more difficulties in correctly interpreting user input. Whereas a car will normally go left if the driver turns the steering wheel in that direction or a vacuum cleaner will start working if one pushes the on-button, interactions between a user and a spoken dialogue system are often hampered by mismatches between the action intended by the user and the action executed by the system. Such mismatches are mainly due to errors in the Automatic Speech Recognition (ASR) and/or the Natural Language Understanding (NLU) component of these systems. To solve these mismatches, users often have to put considerable effort in trying to make it clear to the system that there was a problem, and trying to correct it by re-entering misrecognized or misinterpreted information. Previous research has already brought to light that it is not always easy for users to determine whether their intended actions were carried out correctly or not, in particular when the dialogue system does not give appropriate feedback about its internal representation at the right moment. In addition, users' corrections may miss their goal, because corrections themselves are more difficult for the system to recognize and interpret correctly, which may lead to so-called cyclic (or spiral) errors. That corrections are difficult for ASR systems is generally explained by the fact that they tend to be *hyperarticulated* — higher, louder, longer ... than other turns (Wade et al., 1992; Oviatt et al., 1996; Levow, 1998; Bell and Gustafson, 1999; Shimojima et al., 1999), where ASR models are not well adapted to handle this special speaking style.

The current paper focuses on user corrections, and looks at places where people first become aware of a system problem ("aware sites"). In other papers (Swerts et al., 2000; Hirschberg et al., 2001; Litman et al., 2001), we have already given some descriptive statistics on corrections and aware sites and we have been looking at methods to automatically predict these two utterance categories.

One of our major findings is that prosody, which had already been shown to be a good predictor of misrecognitions (Litman et al., 2000; Hirschberg et al., 2000), is also useful to correctly classify corrections and aware sites. In this paper, we will elaborate more on the exact labeling scheme we used, and add further descriptive statistics. More in particular, we address the question whether there is much variance in the way people react to system errors, and if so, to what extent this variance can be explained on the basis of particular properties of the dialogue system. In the following section we first provide details on the TOOT corpus that we used for our analyses. Then we give information on the labels for corrections and aware sites, and on the actual labeling procedure. The next section gives the results of some descriptive statistics on properties of corrections and aware sites and on their distributions. We will end the paper with a general discussion of our findings.

System Initiative, Explicit Confirmation

T: Which city do you want to go to?
U: Chicago.
S: Do you want to go to Chicago?
U: Yes.

User Initiative, No Confirmation

S: How may I help you?
U: I want to go to Chicago from Baltimore.
S: On which day of the week do you want to leave?
U: I want a train at 8:00.

Mixed Initiative, Implicit Confirmation

S: How may I help you?
U: I want to go to Chicago.
S: I heard you say go to Chicago. Which city do you want to leave from?
U: Baltimore.

Figure 1: Illustrations of various dialogue strategies in TOOT

2 The data

2.1 The TOOT corpus

Our corpus consists of dialogues between human subjects and TOOT, a spoken dialogue system that allows access to train information from the web via telephone. TOOT was collected to study variations in dialogue strategy and in user-adapted interaction (Litman and Pan, 1999). It is implemented using an IVR (interactive voice response) platform developed at AT&T, combining ASR and text-to-speech with a phone interface (Kamm et al., 1997). The system's speech recognizer is a speaker-independent hidden Markov model system with context-dependent phone models for telephone speech and constrained grammars defining vocabulary at any dialogue state. The platform supports barge-in. Subjects performed four tasks with one of several versions of the system that differed in terms of locus of initiative (system, user, or mixed), confirmation strategy (explicit, implicit, or none), and whether these conditions could be changed by the user during the task (adaptive vs. non-adaptive). TOOT's initiative strategy specifies who has control of the dialogue, while TOOT's confirmation strategy specifies how and whether TOOT lets the user know what it just understood. The fragments in Figure 1 provide some illustrations of how dialogues vary with strategy. Subjects were 39 students; 20 native speakers and 19 non-native, 16 female and 23 male. Dialogues were recorded and system and user behavior logged automatically. The *concept accuracy* (CA) of each turn was manually labeled. If the ASR correctly captured all task-related information in the turn (e.g. time, departure and arrival cities), the turn's CA score was 1 (*semantically correct*). Otherwise, the CA score reflected the percentage of correctly recognized task information in the turn. The dialogues were also transcribed and automatically scored in comparison to the ASR recognized string to produce a *word error rate* (WER) for each turn. For the study described below, we examined 2328 user turns (all user input between two system inputs) from 152 dialogues.

2.2 Defining Corrections and Aware Sites

To identify corrections[1] in the corpus two authors independently labeled each turn as to whether or not it constituted a correction of a prior system failure (a rejection or CA error, which were the only system failure subjects were aware of) and subsequently decided upon a consensus label. Note that much of the discrepancies between labels were due to tiredness or incidental sloppiness of individual annotators, rather than true disagreement. Each turn labeled 'correction' was further classified as belonging to one of the following categories: REP (repetition, including repetitions with differences in pronunciation or fluency), PAR (paraphrase), ADD (task-relevant content added, OMIT (content omitted), and ADD/OMIT (content both added and omitted). Repetitions were further divided into repetitions with pronunciation variation (PRON) (e.g. *yes* correcting *yeah*), and repetitions where the correction was pronounced using the same pronunciation as the original turn, but this distinction was difficult to make and turned out not to be useful. User turns which included both corrections and other speech acts were so distinguished by labeling them "2+". For user turns containing a correction plus one or more additional dialogue acts, only the correction is used for purposes of analysis below. We also labeled as *restarts* user corrections which followed non-initial system-initial prompts (e.g. "How may I help you?" or "What city do you want to go to?"); in such cases system and user essentially started the dialogue over from the beginning. Figure 2 shows examples of each correction type and additional label for corrections of system failures on *I want to go to Boston on Sunday.* Note that the utterance on the last line of this figure is labeled 2+PAR, given that this turn consist of two speech acts: the goal of the no-part of this turn is to signal a problem, whereas the remainder of this turn serves to correct a prior error.

Corr Type	Correction
REP	I want to go to Boston on Sunday
PAR	To Boston on Sunday
OMIT	I want to go to Boston
ADD	To Boston on Sunday at 8pm
ADD/ OMIT	I want to arrive Sunday at 8pm
2+PAR	No, to Boston on Sunday

Figure 2: Examples of Correction Types

Each correction was also indexed with an identifier representing the closest prior turn it was correcting, so that we could investigate "chains" of corrections of a single failed turn, by tracing back through subsequent corrections of that turn. Figure 3 shows a fragment of a TOOT dialogue with corrections labeled as discussed above.

We also identified *aware sites* in our corpus — turns where a user, while interacting with a machine, first becomes aware that the system has misrecognized a previous user turn. For our corpus, we tried to determine whether there was some evidence in the user turn indicating that the user had become aware of a mistake in the system's understanding of a previous user turn, and, if so, which previous turn had occasioned that error. Note that such aware sites may or may not also be corrections (another type of post-misrecognition turn), since a user may not immediately provide correcting information. Also, it may take a while before the user is able to notice a system error. Figure 4 shows an example that illustrates cases in which both the user's awareness and correction of a misrecognition occur in the same turn (e.g. turns 1159 and 1160, after system prompts for information already given in turn 1158). It also illustrates cases in which aware sites and corrections occur in different turns. For example, after the immediate explicit system confirmation of turn 1162, the user first becomes aware of the system errors

[1]The labels discussed in this section for corrections and aware sites may well be related to more general dialogue acts, like the ones proposed by (Allen and Core, 1997), but this needs to be explored in more detail in the future.

Turn	Turn ID	Corr of	Corr Type
S: *On which day of the week do you want to leave?*			
S: *Sorry, I didn't hear anything from you. Please tell me your departure date.*			
U: Um tomorrow	776	NA	NA
S: *You can say the departure date. For example, you can say today or tomorrow. Or you can say Monday, Tuesday, Wednesday...*			
U: Tomorrow	777	776	REP
S: *You can say the departure date. For example, you can say today or tomorrow. Or you can say Monday, Tuesday, Wednesday, Thursday, Friday...*			
U: Saturday	778	777	PAR
S: *Do you want to leave on Saturday?*			
U: Yes	779	NA	NA
S: *What time do you want to leave?*			
U: Eight o'clock A M	780	NA	NA
S: *Do you want to leave around 10:00pm?*			
U: Eight o'clock	781	780	OMIT
S: *Do you want to leave around 10:00pm?*			
U: No	782	NA	NA
S: *What time do you want to leave?*			
U: 8:00am	783	781	ADD

Figure 3: Toot Dialogue Fragment with Correction Labels

(turn 1163), then separately corrects them (turn 1164). When no immediate confirmation of an utterance occurs (as with turn 1158), it may take several turns before the user becomes aware of any misrecognition errors. For example, it is not until turn 1161 that the user first becomes aware of the error in date and time from 1158; the user then corrects the error in 1162. Of all turns in our corpus, 13% represent cases of turns that are only corrections, 14% are only aware sites, and 16% are turns where aware sites and corrections co-occur. Also, note that turns 1162 and 1164 in this dialogue fragment represent cases of restarts after a system-initial, non-initial prompt ("How may I help you?").

3 Characteristics of corrections and aware sites in TOOT

3.1 Corrections

We identified 29% of the turns in the corpus as corrections. These were a serious problem for the system, being misrecognized in terms of transcription error about twice as often (61%) as non-corrections (33%) and in terms of concept error in 53% of cases for corrections compared with 21% concept error for non-corrections.

Corrections differed from non-corrections in other ways as well, which are suggestive of how they might be identified by an SDS. We have previously discussed the prosodic and acoustic differences between corrections and non-correction (Swerts et al., 2000; Hirschberg et al., 2001). In particular, in line with previous studies, we found that corrections more often tend to be hyperarticulated (longer, louder, with great pitch excursions,...) than other turns. More recently we have also discovered differences in other dimensions.

First, corrections appear to be longer than non-corrections in numbers of words, when we examine actual human transcriptions of

Turn		Turn ID	Aware of	Corr of
S:	*How may I help you?*			
U:	Get me the train to New York City from Baltimore on Sunday at eight thirty pm	1158	NA	NA
S:	*Which city do you want to go to?*			
U:	New York City	1159	1158	1158
S:	*Which city do you want to leave from?*			
U:	Baltimore	1160	1158	1158
S:	*Do you want me to find the trains from Baltimore to New York City today at anytime now?*			
U:	No	1161	1158	NA
S:	*How may I help you?*			
U:	Get me the train to New York City from Baltimore on Sunday at eight thirty pm	1162	NA	1158
S:	*Do you want me to find the trains from Baltimore to New York City today at anytime now?*			
U:	No	1163	1162	NA
S:	*How may I help you?*			
U:	Get me the train on Sunday at eight thirty pm	1164	NA	1162

Figure 4: Dialogue Fragment with Aware and Correction Labels.

them, both in absolute terms (T=17.68; df=2326; p=0) and also controlling for speaker (T=5.32; df=38; p=0). Even the ASR hypotheses show this difference, with hypotheses of corrections being longer in absolute terms (T=13.72; df=2326; p=0) and across speakers (T=5.18; df=38; p=0).

Of the correction types we labeled, the largest number were REPs and OMITs, as shown in Table 1, which shows over-all distribution of correction types, and distributions for each type of system failure corrected. Table 1 shows that 39% of TOOT corrections were simple repetitions of the previously misrecognized turn. While this strategy is often suboptimal in correcting ASR errors (Levow, 1998), REPs (45% error) and OMITs (52% error) were better recognized than ADDs (90% error) and PARs (72% error). Thus, overall, users tend to have a preference for correction types that are more likely to be successful. That REPs and OMITs are more often correctly recognized can be linked to the observation that they tend to be realized with prosody which is less marked than the prosody on ADDs and PARs. Table 2 shows that REPs and OMITs are closer to normal utterances in terms of their prosodic features than ADDs, which are considerably higher, longer and slower. This is in line with our previous observations that marked settings for these prosodic features more often lead to recognition errors.

What the user was correcting also influenced the type of correction chosen. Table 1 shows that corrections of misrecognitions (Post-Mrec) were more likely to omit information present in the original turn (OMITs), while corrections of rejections (Post-Rej) were more likely to be simple repetitions. The latter finding is not surprising, since the rejection message for tasks was always a close paraphrase of "Sorry, I can't understand you. Can you please repeat your utterance?" However, it does suggest the surprising power of system directions, and how important it is to craft prompts to favor the type of correction most easily recognized by the system.

Corrections following system restarts differed in type somewhat from other corrections, with more turns adding new material to the correction and fewer of them repeating

	ADD	ADD/OMIT	OMIT	PAR	REP
All	8%	2%	32%	19%	39%
% Mrec(WER)	90%	93%	52%	72%	45%
% Mrec(CA)	88%	71%	47%	65%	45%
Post-Mrec	7%	3%	40%	18%	32%
Post-Rej	6%	0%	7%	28%	59%

Table 1: Distribution of Correction Types

Feature	Normal	ADD	ADD/OMIT	OMIT	PAR	REP
F0max (Hz)	219.4	286.3	252.9	236.7	252.1	239.9
rmsmax	1495.0	1868.1	2646.3	1698.0	1852.4	2024.6
dur (s)	1.4	6.8	4.1	2.3	4.7	2.5
tempo (sylls/s)	2.5	1.7	1.6	2.9	2.1	2.3

Table 2: Averages for different prosodic features of different Correction Types

the original turn.

Dialogue strategy clearly affected the type of correction users made. For example, users more frequently repeat their misrecognized utterance in the SystemExplicit condition, than in the MixedImplicit or UserNoConfirm; the latter conditions have larger proportions of OMITs and ADDs. This is an important observation given that this suggests that some dialogue strategies lead to correction types, such as ADDs, which are more likely to be misrecognized than correction types elicited by other strategies.

As noted above, corrections in the TOOT corpus often take the form of chains of corrections of a single original error. Looking back at Figure 3, for example, we see two chains of corrections: In the first, which begins with the misrecognition of turn 776 ("Um, tomorrow"), the user repeats the original phrase and then provides a paraphrase ("Saturday"), which is correctly recognized. In the second, beginning with turn 780, the time of departure is misrecognized. The user omits some information ("am") in turn 781, but without success; an ADD correction follows, with the previously omitted information restored, in turn 783. Elsewhere (Swerts et al. 2000), we have shown that chain position has an influence on correction behaviour in the sense that more distant corrections tend to be misrecognized more often than corrections closer to the original error.

3.2 Aware Sites

708 (30%) of the turns in our corpus were labeled aware sites. The majority of these turns (89%) immediately follow the system failures they react to, unlike the more complex cases in Figure 4 above. If a system would be able to detect aware sites with a reasonable accuracy, this would be useful, given that the system would then be able to correctly guess in the majority of the cases that the problem occurred in the preceding turn. Aware turns, like corrections, tend to be misrecognized at a higher rate than other turns; in terms of transcription accuracy, 50% of awares are misrecognized vs. 35% of other turns, and in terms of concept accuracy, 39% of awares are misrecognized compared to 27% of other turns. In other words, both types of post-error utterances, i.e., corrections and aware sites, share the fact that they tend to lead to additional errors. But whereas we have shown above that for corrections this is probably caused by the fact that these utterances are uttered in a hyperarticulated speaking style, we do not find differences in hyperarticulation between aware sites and 'normal utterances' (T= 0.9085; df=38; p=0.3693). This could mean that these sites are realized in a speaking style which is not perceptibly different from normal speaking style

	ADD	ADD/OMIT	OMIT	PAR	REP
MixedExplicit	1	0	4	1	4
MixedImplicit	16	8	58	44	64
MixedNoConfirm	0	0	2	0	1
SystemExplicit	2	2	8	31	67
SystemImplicit	0	1	18	0	20
SystemNoConfirm	0	0	5	0	4
UserExplicit	0	0	0	1	1
UserImplicit	1	0	4	3	6
UserNoConfirm	31	3	116	47	98

Table 3: Number of Correction Types for different dialogue strategies

	Single no	Other Turns
Aware site	162	546
Not Aware site	122	1498

Table 4: Distribution of single no utterances and other turns for aware sites versus other utterances

when judged by human labelers, but which is still sufficiently different to cause problems for an ASR system.

In terms of distinguishing features which might explain or help to identify these turns, we have previously examined the acoustic and prosodic features of aware sites (Litman et al., 2001). Here we present some additional features. Aware sites appear to be significantly shorter, in general, than other turns, both in absolute terms and controlling for speaker variation, and whether we examine the ASR transcription (absolute: T=4.86; df=2326; p=0; speaker-controlled: T=5.37; df=38; p=0) or the human one (absolute: T=3.45; df=2326; p<.0001; speaker-controlled: T=4.69; df=38; p=0). A sizable but not overwhelming number of aware sites in fact consist of a simple negation (i.e., a variant of the word 'no') (see Table 4). This at the same time shows that a simple no-detector will not be sufficient as an indicator of aware sites (see also (Krahmer et al., 1999; Krahmer et al., to appear)), given that most aware sites are more complex than that, such as turns 1159 and 1160 in the example of Figure 4. More concretely, Table 4 shows that a single no would correctly predict that the turn is an aware site with a precision of only 57% and a recall of only 23%.

4 Discussion

This paper has dealt with user corrections and aware sites of system errors in the TOOT spoken dialogue system. We have described our corpus, and have given details on our procedure to label corrections and aware sites. Then, we have shown that corrections and aware sites exhibit some prosodic and other properties which set them apart from 'normal' utterances. It appears that some correction types, such as simple repeats, are more likely to be correctly recognized than other types, such as paraphrases. We have also presented evidence that system dialogue strategy affects users' choice of correction type, suggesting that strategy-specific methods of detecting or coaching users on corrections may be useful. Aware sites tend to be shorter than other utterances, and are also more difficult to recognize correctly for the ASR system.

In addition to the descriptive study presented in this paper, we have also tried to automatically predict corrections and aware sites using the machine learning program RIPPER (Cohen, 1996). These experiments show that corrections and aware sites can be classified as such automatically, with a considerable degree of accuracy (Litman et al., 2001; Hirschberg et al., 2001). Such classification, we believe, will be especially useful in error-handling for SDS. If aware sites are detectable, they can function as backward-looking

error-signaling devices, making it clear to the system that something has gone wrong in the preceding context, so that, for example, the system can reprompt for information. In this way, they are similar to what others have termed 'go-back' signals (Krahmer et al., 1999). Aware sites can also be used as forward-looking signals, indicating upcoming corrections or more drastic changes in user behavior, such as complete restarts of the task. Given that, in current systems, both corrections and restarts often lead to recognition error (Swerts et al., 2000), aware sites may be useful in preparing systems to deal with such problems. An accurate detection of turns that are corrections may trigger the use of specially trained ASR models to better recognize corrections, or can be used to change dialogue strategy (e.g. from user or mixed initiative to system initiative after errors).

References

J. Allen and M. Core. 1997. Dialogue markup in several layers. Draft contribution for the Discourse Resource Initiative.

L. Bell and J. Gustafson. 1999. Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. In *Proceedings of ICPhS-99*, San Francisco. International Congress of Phonetic Sciences.

W. Cohen. 1996. Learning trees and rules with set-valued features. In *14th Conference of the American Association of Artificial Intelligence, AAAI*.

J. Hirschberg, D. Litman, and M. Swerts. 2000. Generalizing prosodic prediction of speech recognition errors. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing.

J. Hirschberg, D. Litman, and M. Swerts. 2001. Identifying user corrections automatically in spoken dialogue systems. In *Proceedings of NAACL-2001*, Pittsburgh.

C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. 1997. Evaluating spoken dialogue systems for telecommunication services. In *Proc. EUROSPEECH-97*, Rhodes.

E. Krahmer, M. Swerts, M. Theune, and M. Weegels. 1999. Error spotting in human-machine interactions. In *Proceedings of EUROSPEECH-99*.

E. Krahmer, M. Swerts, M. Theune, and M. Weegels. to appear. The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. Accepted for *Speech Communication*.

G. Levow. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics, COLING/ACL 98*, pages 736–742.

D. Litman, J. Hirschberg, and M. Swerts. 2000. Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of NAACL-00*, Seattle, May.

D. Litman, J. Hirschberg, and M. Swerts. 2001. Predicting User Reactions to System Error. In *Proceedings of ACL-01*, Toulouse, July.

D. Litman and S. Pan. 1999. Empirically evaluating an adaptable spoken dialogue system. In *Proceedings tth International conference on User Modeling*.

S. L. Oviatt, G. Levow, M. MacEarchern, and K. Kuhn. 1996. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of ICSLP-96*, pages 801–804, Philadelphia.

A. Shimojima, K. Katagiri, H. Koiso, and M. Swerts. 1999. An experimental study on the informational and grounding functions of prosodic features of Japanese echoic responses. In *Proceedings of the ESCA Workshop on Dialogue and Prosody*, pages 187–192, Veldhoven.

M. Swerts, D. Litman, and J. Hirschberg. 2000. Corrections in spoken dialogue systems. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing.

E. Wade, E. E. Shriberg, and P. J. Price. 1992. User behaviors affecting speech recognition. In *Proceedings of ICSLP-92*, volume 2, pages 995–998, Banff.