Effects of Speech Recognition Accuracy on the Performance of DARPA Communicator Spoken Dialogue Systems

Gregory A. Sanders and Audrey N. Le

National Institute of Standards and Technology

Author to be contacted for correspondence:

Gregory Sanders
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899 – 8940

(301) 975-4451 voice
(301) 670-0939 FAX
gregory.sanders@nist.gov

Abstract:

The DARPA Communicator program explored ways to construct better spoken-dialogue systems, with which users interact via speech alone to perform relatively complex tasks such as travel planning. During 2000 and 2001 two large data sets were collected from sessions in which paid users did travel planning using the Communicator systems that had been built by eight research groups. The research groups improved their systems intensively during the ten months between the two data collections. In this paper, we analyze these data sets to estimate the effects of speech recognition accuracy, as measured by Word Error Rate (WER), on other metrics. The effects that we found were linear. We found correlation between WER and Task Completion, and that correlation, unexpectedly, remained more or less linear even for high values of WER. The picture for User Satisfaction metrics is more complex: we found little effect of WER on User Satisfaction for WER less than about 35% to 40% in the 2001 data. The size of the effect of WER on Task Completion was less in 2001 than in 2000, and we believe this difference is due to improved strategies for accomplishing tasks despite speech recognition errors, which is an important accomplishment of the research groups who built the Communicator implementations. We show that additional factors must account for much of the variability in task success, and we present multivariate linear regression models for task success on the 2001 data. We also discuss the apparent gaps in the coverage of our metrics for spoken dialogue systems.

Keywords:

Communicator, spoken dialogue, word error rate, task success, completion, efficiency, user satisfaction

Effects of Speech Recognition Accuracy on the Performance of DARPA Communicator Spoken Dialogue Systems

The DARPA Communicator program focused on research into ways to construct better spoken-dialogue systems, which users interact with via speech alone, to perform relatively complex tasks such as travel planning. Although not discussed in this paper, the program also created toolkits and architectures for faster, easier, and cheaper creation of high quality interactive spoken dialogue systems. These toolkits and architectures make it much easier for first-time developers to create such systems.

The data analyzed in this paper come from eight different Communicator systems, built by eight different research groups. A research team at MIT designed an architecture, called Galaxy-II (Seneff, et al., 1998; Polifroni and Seneff, 2000), that was adapted in collaboration with a research team at the MITRE Corporation and used by all Communicator implementations.

The program included collection of two large-scale data sets, in 2000 and 2001. In both data collections, users made travel arrangements using the Communicator systems. All interaction between the users and systems was spoken conversations via telephone. We do note that relatively few disfluencies (Shriberg, 1994; Oviatt, 1995) arose in the user speech in these Communicator dialogues. The interaction was logged in detail (Aberdeen, 2000); the user utterances were transcribed by humans (Sanders and ATIS Committee, 2001); and metrics were calculated from these logs (Le, Sanders, and Aberdeen, 2001) and transcriptions. We analyzed the data, trying to uncover the factors that determine the performance of spoken dialogue systems.

We compare the 2000 data with the 2001 data, exploring the effects of Automatic Speech Recognition (ASR) accuracy on three measures of system performance: task success, user satisfaction, and efficiency of performing the task. Task success receives the most intensive analysis here.

ASR accuracy is usually measured by Word Error Rate (WER), defined as the number of errors divided by the number of words spoken. The number of errors is the sum of the number of words inserted (e.g., noise interpreted as words), deleted, or substituted. For example, if the user said, “I want to recognize speech,” but the system heard “I want to wreck a nice beach”, we would score two substitutions and two insertions (four errors) divided by five words actually spoken, giving WER = 80%. WER can exceed 100% if the ASR system inserts enough words.

In related work in a much different data retrieval application, Bonneau-Maynard, Devillers, and Rosset (2000) also investigated the effects of ASR performance, among other factors, although they focused on user satisfaction. And previous work by Walker et al. (2001, 2002) gives an intensive analysis of user satisfaction in the Communicator systems.

Strong ASR is very helpful but is not the whole explanation for success: indeed, some systems with only moderately good ASR performance did well on crucial metrics such as task completion and user satisfaction. In this paper, we try to answer three questions about the effects of ASR performance:

- How high can the Word Err Rate (WER) be and still have a system do well?
- Are the effects of WER linear?
- Does performance fall off a cliff above some WER value?

Our brief conclusions about these three questions are as follows. Most calls completed the task when WER was 50% or less (WER was 50% or less on over 90% of the calls in 2001). Averaging across systems, ASR accuracy appears to have a linear correlation with successful completion of air-travel planning tasks, although we present a multivariate linear regression showing that there are other important factors. We expected performance to deteriorate sharply above some level of WER, but this does not appear to be the case; efficiency as measured by time and user words on task is, however, more variable above a WER of 35% to 40%.

Earlier analyses by Walker, et al., (2001, 2002) constructed a multivariate linear-regression model for user satisfaction, characterizing the overall performance of different Communicator implementations in the data collection that was done during June 2000. Our short conclusion about user satisfaction is that there are *several* important determinants of user satisfaction. In particular, speech recognition accuracy alone accounts for little of the variation in user satisfaction, whereas multivariate linear-regression approaches give an explanatory model that accounts for much more of the variation in user satisfaction. The interested reader is referred to the Walker et al. (2001, 2002) papers. In this paper, we include multivariate models for task completion, similar to the models for user satisfaction that were given by Walker et al.

The User Questionnaire

In both 2000 and 2001, immediately after each call the user filled out a questionnaire giving opinions about the call that was just completed. The questionnaire consisted of one yes/no question and five Likert-style items. The yes/no question asked, “Were you able to complete your entire task?” We call this item *Perceived Task Completion* (*PTC*).

Each Likert-style item consisted of a declarative statement for which the user was to choose one of five possible responses: “completely agree”, “agree”, “neutral”, “disagree, or “completely disagree”. The five Likert-style items were as follows. Note each is worded positively, so that Completely Agree is the most favorable response.

Would Use Regularly (WUR)
“Based on my experience in this conversation using this system to get travel information, I would like to use this system regularly.”

Expected Behavior
“The system worked the way I expected it to in this conversation.”

Interface Confidence
“In this conversation, I knew what I could say or do at each point in the dialogue.”

Task Ease
“In this conversation, it was easy to get the information that I wanted.”

TTS Performance
“I found it easy to understand what the system said.”

Table 1. The Likert-style items from the user questionnaire.

These five Likert-style items constitute (collectively) a user questionnaire that had been used previously by AT&T speech researchers to get user feedback on factors that the researchers believed to be relevant to user satisfaction or to system success. For purposes of data analysis, we coded the five possible responses to a Likert item as 5.0 for “completely agree” through 1.0 for “completely disagree”, with 3.0 being neutral. Our analysis suggests that this coding as equally-spaced numeric values may be reasonable, but that perhaps Completely Disagree should be given a numeric value of 0 rather than 1. For example, mean ATC for each of the possible responses to WUR is 0.61, 0.74, 0.81, 0.88, and 0.93 — with the means for the two adjoining responses Completely Disagree (0.74) and Disagree (0.81) differing about twice as much as the means for any other two adjoining responses.

The *TTS Performance* item on the questionnaire was apparently intended to get the user’s opinion of the text-to-speech (TTS) synthesis. The responses on that item, however, correlate with other items on the user questionnaire: in particular, multivariate linear regression shows that the *Task Ease* and *Interface Confidence* items together account for over 46% of the variance in the TTS Performance item. If the responses to the *TTS Performance* item actually reflected the performance of the text-to-speech synthesis, then those responses should be about the same for all calls made by a given user in 2001. But the responses were not constant, so it appears users did not know how to interpret the *TTS Performance* item. We do not know how to interpret the responses to this item and have not analyzed them.

Description of the 2000 data collection

Nine Communicator systems participated in the 2000 data collection (eight of these also participated in the 2001 data collection). In that 2000 data collection, we intended to have 72 subjects each make nine calls, one to each system. The order in which subjects encountered the systems was counter-balanced. For each of the first seven calls, we provided an itinerary for the subject to plan. We created seven scenario templates to be used for these first seven calls. For the eighth and ninth calls we asked the subjects to plan real trips they might want to take. Subjects were given several days during a three-week period to make their calls. We ended up with 81 paid subjects who made 533 analyzable calls for which we also got analyzable user questionnaires and which are included in this paper. There were an additional 40 calls from the site that did not participate in 2001 and they are therefore omitted from the results in this paper, in order to increase comparability between the two years' data.

A typical example of a template called for a round-trip flight between a large domestic airport and a large foreign airport, using a specified airline. Each instantiation of that template would specify the particular airports, dates, times of day, airline, etc. The templates which had the lowest performance involved one or more small/remote domestic airports. For each day that the data collection was going on, we created an instantiation for each of the seven templates, by choosing the specific cities/airports, dates, and airline preferences that the users were to request from the system. All instantiations of the same template were supposed to be of equivalent difficulty, and we did not find statistically significant evidence to the contrary, although the number of calls per instantiation may have been too low to do so. Results on the last two calls (planning real trips) turned out to be similar to results on the first seven calls (planning hypothetical trips defined by the scenarios).

Description of the 2001 data collection

The 2001 data collection began about ten months after the 2000 data collection and was organized differently. As a result, most of the 2001 tasks are not directly comparable to those in the 2000 collection. As was mentioned, each of the users in 2000 was to call all the systems (one call to each system). So each questionnaire response in 2000 dealt with a single call to a single system, and the call was always the user's first call to that system. In contrast, in the

2001 collection each user was assigned to one system and used it repeatedly (a within-system rather than within-subject design). Although (as in 2000) each 2001 questionnaire response deals with one call, most of the 2001 calls are not the user's first call to the system. Assignment of users to systems was systematically random, with more or less equal numbers of users assigned to each system. The 2001 collection included some hypothetical travel scenarios but had nothing corresponding to multiple instantiations of a scenario template.

We recruited approximately 290 frequent travelers to use the systems during the 2001 data collection. They were divided into two groups. Each user in the first group (which we refer to as the "short subjects") was asked to make four calls, planning actual intended air-travel in each call. Seventy-seven of the short subjects completed four analyzable calls. Each user in the second group (which we refer to as the "long subjects") was to make ten calls, doing hypothetical itineraries in the first four and last two calls and planning actual intended air travel in the other four calls (the fifth through eighth calls). Sixty-nine of these long subjects completed ten analyzable calls. We used the calls from all the callers, whether or not they completed all their scheduled calls. In total, we ended up with 290 paid subjects in 2001 who made just over 1200 analyzable calls.

The first two of the 2001 hypothetical scenarios were comparable to 2000 scenario instantiations. And the four calls made by the short subjects (planning real trips) can be compared to the fifth through eighth calls made by the long subjects (planning real trips) — but with the long subjects having more experience by the time they made those calls, so as to give a way to look at learning effects. These calls can also be compared to the eighth and ninth calls made by subjects in 2000. The data do suggest that most systems improved between 2000 and 2001.

As in 2000, the subjects filled out a brief user questionnaire after each call, which included the same five Likert-style items as in 2000. When subjects answered the Perceived Task Completion (PTC) question on the user questionnaire, we wanted them to tell us whether they were able to find air-travel reservations that matched what they had requested, even if they were not able to find desired hotel and/or rental cars, but the initial instructions about the PTC item given to the users at the beginning of the 2001 data collection did not make that clear. We know from user inquiries that users were confused. We fixed this problem with the instructions about July 1, 2001. We do not trust the user's PTC responses before that fix, so to maximize comparability of 2001 PTC data with the 2000 PTC results, this paper employs a composite PTC metric for the 2001 data: using Task Completion as annotated by the sites (ATC) for calls made before July 1 and using the PTC data from the user questionnaires for calls made after that date. We believe the 2001 PTC (really Composite-PTC) values are reasonably comparable to the 2000 PTC results. Substituting ATC for PTC for calls made before July 1 reverses the PTC result for 48 calls, which is less than 4% of the calls. Thus, we had basically the same user questionnaire in 2001 as in 2000, and the on-line presentation of the questionnnaire in 2001 was likewise the same as in 2000. The resulting data from the questionnaires should be comparable between the two years.

What data did we choose to analyze in this paper?

One overarching goal of the analysis presented in this paper was to compare the June 2000 results to the results on the 2001 data collection (running April through September 2001). Our analysis of the 2000 data omits all calls that were handled by the system that did not participate in 2001, to increase comparability between the two years' data.

In our analysis of task completion, we wanted to look at how strongly ASR accuracy correlated with task completion. Over a group of calls, task completion (ATC or PTC) is the fraction (0.0 through 1.0) of the calls that completed the task.

We also wanted to look at how strongly user satisfaction correlated with WER. This raised the question of how best to measure user satisfaction for this purpose. The PARADISE framework (Walker, Litman, Kamm, and Abella, 1997; Walker, Kamm, and Litman, 2000), has been used in previous analyses of the Communicator data by Walker, et. al. (2001). Those analyses of the Communicator data used the average of the five Likert-style items (in Table 1) as the measure of user satisfaction. However, some of the Likert-style items do not ask about user satisfaction per se, and as has been mentioned we do not know how to interpret the responses to the *TTS Performance* item on the questionnaire. So, for our analysis focusing on the effects of ASR it seemed more appropriate to measure user satisfaction with just the Likert-style item, WUR.

We did consider some similar alternative metrics for user satisfaction. Walker et al. (2001) presented a linear regression showing that the variance on the WUR item is strongly correlated with the variance in the sum of all five items. Doing pair-wise linear regressions between WUR and the other items on the user questionnaire reveals that the correlation is strongest (r-square = 0.54) between WUR and the *Task Ease* item. We could average WUR and *Task Ease* as our metric for user satisfaction, but we feel that would over-weight ease of use. There is also a strong correlation between *Task Ease* item and the *Expected Behavior* item that probably reflects whether the system matched the user's preconceptions or behaved consistently from one call to the next. We believe the characteristics reflected in the *Expected Behavior* item do not result from user satisfaction but rather tend to produce it. Therefore, as mentioned, we used WUR alone as our user satisfaction metric.

We turn now to discussing why some of our intended analyses could not be performed due to insufficient data. Two kinds of problems with this data collection made statistical analysis more difficult. First, in 2000 due to problems scheduling users, some scenario template instantiations were used in fewer than four calls (a total of 29 calls in such instantiations). Metrics based on so few calls are relatively uninformative, so we filtered out those instantiations. Second, in both years some calls ended with a system crash or other problem before many words had yet been spoken, and as a result too few words were spoken to have a meaningful measure of WER. Ultimately, we decided to filter out calls that were shorter (fewer words and turns) than the shortest calls that completed a travel planning task, which eliminated about 5% of the calls as too short to measure WER meaningfully. Ignoring one outlier call, the shortest calls in the entire 2000 data set that completed a travel planning task had 6 user sentences (utterances) totaling 16 user words in the hand transcription of the data. We kept the 2000 calls that were at least that long. In the 2001 data, the shortest successful call had 5 user turns and 10 user words, so we kept 2001 calls that were at least that long.

Both these problems (too few calls per template instantiation, too few words and turns per call) are unrelated to the metrics that we analyzed. We filtered out the calls that are in these insufficient-data categories, in order to improve our confidence in the statistical results. We have done most of our analyses both with and without this filtering-out of calls and scenarios for which there was very little data. We believe the filtering has made the numbers in the analysis more statistically meaningful (by analyzing only statistics that are based on sufficient data). But filtering the data pool in these ways has changed neither the patterns identified nor the qualitative conclusions drawn, and there are no obvious patterns of increasing or decreasing the numeric values in any aggregated results.

Logfile Defects

In total, 1.3% of the 2000 calls and over 12% of the 2001 calls had logfile defects, missing questionnaires, or other problems that made it impossible to determine the value of some metric(s), and we excluded those calls from the results reported in this paper (we want all metrics to be based on the same set of calls so that correlations between the metrics will be more meaningful). An additional 4.1% of the 2000 logfiles that are included in the analyses in this paper had defects that would prevent certain analyses not reported on here.

Twelve percent is a significant fraction of the 2001 data. We have identified needed changes in the logfile format to directly associate utterance begin and end times with utterance texts and we note that we needed the systems to log begin and end times even for utterances that were barged-in upon and abandoned, as those times cannot be generated by post-hoc analysis of the existing logfiles. These changes would have eliminated most of the data loss in 2001.

Metrics for the Performance of Automatic Speech Recognition

We turn now to the actual analyses. One hypothesis we investigated in the 2000 data was the idea that the important ASR errors are those on names of airlines, days of the week, months, spoken numbers, and similar “keywords”. Intuitively, these errors on “keywords” would be expected to be more important. But we have been *unable* to demonstrate that any metric for ASR performance on just keywords (compared to ASR performance on all words) has any advantage at predicting task success or user satisfaction.

On the 2000 data, we had four metrics for the performance of automatic speech recognition (ASR): the traditional Word Error Rate (WER) over all words, a Keyword-weighted Word Error Rate (KW_WER), the Sentence Error Rate (SER), and a Keyword-weighted Sentence Error Rate (KW_SER). A “sentence” is really a complete user utterance in the Communicator data. The definition of SER and KW_SER is that the sentence (utterance) is wrong if there is any ASR error in it, so that word errors determine sentence errors. We wanted to know whether anything is gained by using all four of these ASR metrics, a question to which we now turn.

What appears to be the case is that misrecognitions of keywords do trigger corrections (and user dissatisfaction). But successful repairs of errors are more likely if the system’s ASR is accurate (low WER) on *all* words. Users do try to repair the important misrecognitions of keywords needed for successful task performance. Note that users will be highly motivated to repair important errors such as dates and city names if they want to plan actual travel. Note that even human-to-human dialogues have misunderstandings that must be repaired.

Simple linear regression analyses on the four ASR metrics shows the four ASR metrics are strongly correlated, as one would expect. Figure 1 shows that very few calls in 2000 had WER > 85%. A simple linear regression on the 2000 data for all calls that had WER < 85%, with WER as the predictor value and SER as the response variable, shows WER and SER to be correlated, with r-square = 0.65, meaning that 65% of the variation in the SER was correlated with the variation in WER. The relationship between the definition of SER and the definition of WER actually entails that they are correlated, so we attribute the variation in SER to the variation in WER. The slope of the least-squares regression line for WER predicting SER over this data is 0.94, suggesting that on average SER is 0.94 * WER. The p-value for the slope of the regression line is less than 0.0001, meaning it is highly unlikely that the slope is significantly positive (the 95% confidence interval for this slope is 0.89 through 1.00). Thus, we conclude that SER gives somewhat the same information as WER.

Over exactly the same 2000 data set, the corresponding regression with WER as the predictor value and KW_WER as the response variable gives r-square 0.91 — suggesting that about 91% of the variation in KW_WER can be accounted for by the variation in WER, or vice-versa. The slope of the least-squares regression line is 1.38, and the p-value for the slope is effectively zero, which means we are sure the slope is significantly positive. And in fact, the 95% confidence interval for this slope is 1.34 through 1.41. These r-square and slope values suggest that WER and KW_WER give almost the same information, which we find plausible. Similar comments can be made about the relationship between SER and KW_SER. Also, the two word error rate metrics consistently give stronger correlations with task success and with user satisfaction than do the two sentence error rate metrics.

Aggregating the data in various ways, it transpired that on average the correlations of the ASR metrics with task success and with user satisfaction are stronger (higher r-square values) when using the traditional WER metric as the predictor variable than when using any of the other three ASR metrics. So the traditional WER metric seems sufficient, and we will present ASR results using WER alone.

Why We Did Not Pool the 2000 Data With The 2001 Data

If we could legitimately pool the data from these two data collections then the analysis would gain statistical power (ability to detect an effect when one is present). Pooling the data makes sense only if the data from the two years is comparable. But in important respects, the data from the two years seems to have different characteristics. The mean WER improved (was lower) for most of the systems by 2001 and its standard deviation correspondingly decreased. In fact, pooling the data from the seven Communicator systems together, the *z* statistic for the difference in the mean WER between the two years is 3.7, so we are confident that mean WER was less in 2001. In 2000, the call with the lowest WER that was greater than 0 had WER = 2.0%, and there were few calls in 2000 with WER > 55%. That range (2.0% ≤ WER ≤ 55%) is well represented by calls in both years' data sets. We compared the 2000 and 2001 subsets of calls that are in that range in order to decide whether to pool the data from 2000 with the data from 2001.

The slope of the least-squares regression lines, for WER predicting Perceived Task Completion (PTC) on these subsets differs between the two years — it was -0.0088 on the 2000 data and -0.0056 on the 2001 data, a difference that is statistically significant at the 95% confidence level. And mean task completion as measured by ATC improved significantly between the two years' subsets ($z > 5.7$; $p \approx 0.0$), as did PTC ($z = 3.5$; $p < .001$). However the slope for WER predicting ATC on this range was almost identical in 2000 (-0.0088) and 2001 (-0.0081). Finally, on this subset of the data, there were more or less equivalent responses on the WUR user questionnaire item (see Figure 5). Noting the improvements between the two years in WER, PTC, and ATC, particularly the evidence of better PTC on calls with high error rates in 2001, we concluded the two years' data should not be pooled.

Basic Characteristics of the Data

We begin with a graphical look at the data. As can be seen in the two histograms with 5% bin-width (Figures 1 and 2) there were more calls in 2001 than in 2000. In Figures 1 and 2, one can see that an obvious constraint on data analysis is that there is little data for high WER values. The leftmost bar in these histograms shows the calls with WER = 0.0%, and the other bars show a range of values (e.g., the second bar shows $0.01\% \leq \text{WER} \leq 5.0\%$). In both years, there were at least 16 calls per bin through 55% WER, which is over 93% of the calls in both years.

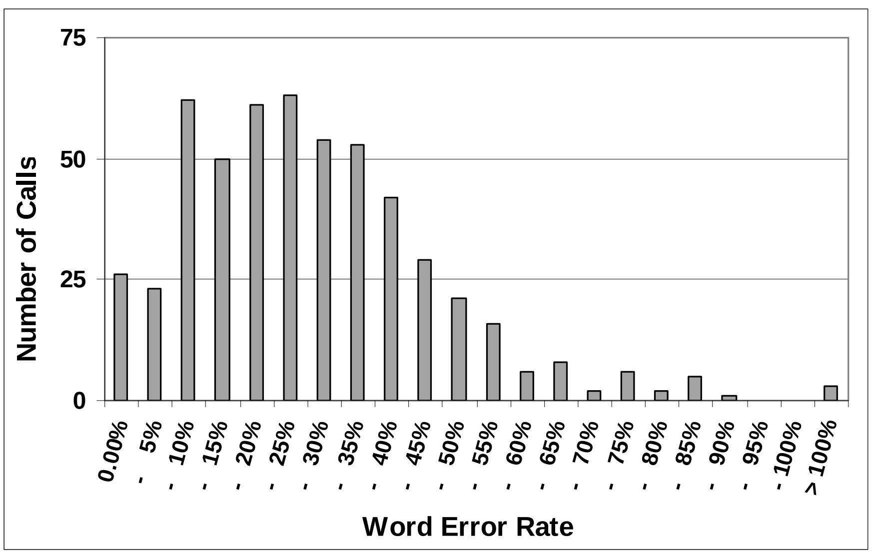

Figure 1. Histogram of calls by WER for 2000.

Mean WER was 26.39% in 2000 (std. dev. 19.14) and 22.54% in 2001 (std. dev. 21.30), which is a significant difference in mean WER ($z = 3.7$, $p < .001$). One can see in Figures 1 and 2 that the distribution of WER is skewed toward high values of WER and that there are outlier calls with very high WER in both years.

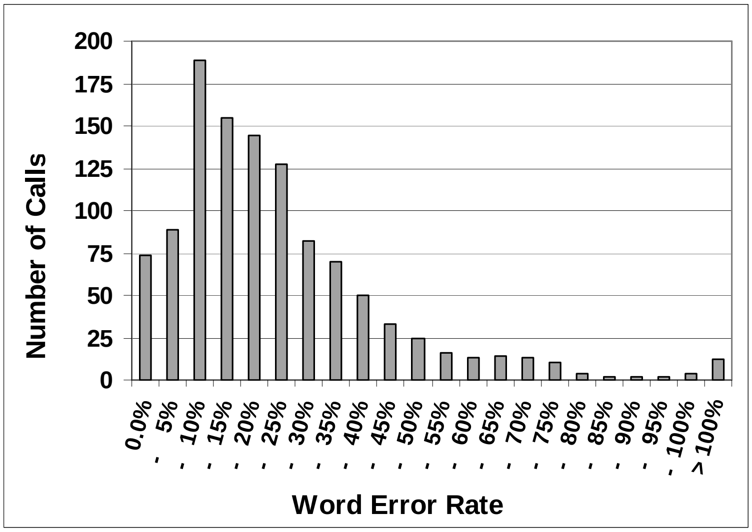

Figure 2. Histogram of calls by WER for 2001.

In 2001 (Figure 2) there are at least 13 calls per bin through 70% WER despite the lower overall WER (because there were more calls in 2001). We note that in 2001 80% of the calls had WER less than 35%, 90% had WER less than 46.5% and 95% had WER less than 65%. Later figures in this paper will present mean values of various metrics for these histogram bins — through 55% WER for 2000 and through 70% WER for 2001. In the case of smaller effect sizes, the reader may wish to keep in mind that there are fewer calls per bin above 45%.

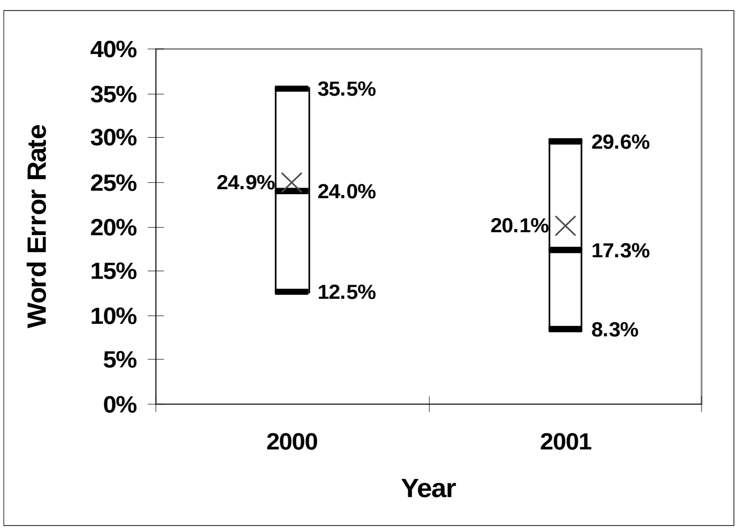

Figure 3. Quartiles and 90% Trimmed Mean of WER for 2000 and 2001, showing improvement in ASR.

Figure 3, gives an alternate look at the data eliminating the outliers and most of the skew. It shows that the median WER (middle line each bar) was 24.0% in 2000 and 17.3% in 2001 and shows that half the calls in 2000 had WER in the range 12.5% through 35.5%, where half the calls in 2001 had WER in the range 8.3% through 29.6%.

The mean over the middle 90% of the data is also shown (the X in the bars). These improvements in WER are statistically significant — e.g., for the difference between the trimmed means ($z = 6.5$; $p \approx 0$). Improvement in WER can be seen in the data for most of the Communicator sites/implementations individually as well.

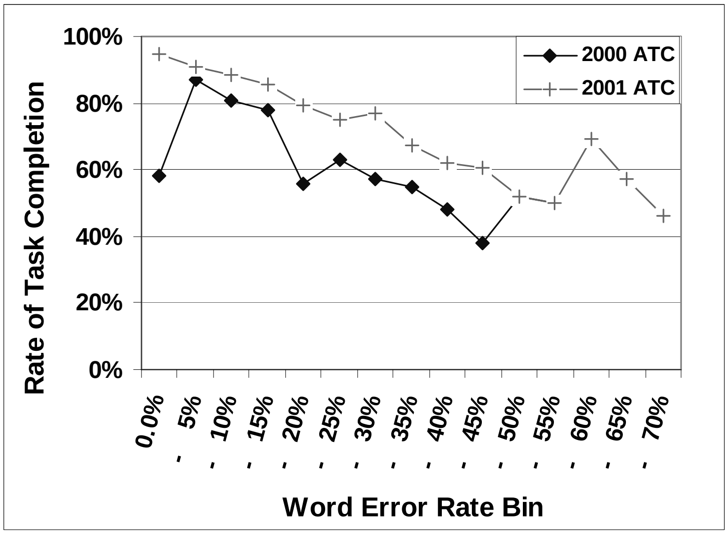

Figure 4. Annotated Task Completion (ATC) data for 2000 and 2001.

As has been mentioned, the difference between mean ATC for 2000 and mean ATC for 2001 for calls aggregated over the whole range 2.0% ≤ WER ≤ 55% is highly significant. Figure 4 shows ATC differences between the two years by WER histogram bin (the bins correspond to Figures 1 and 2). At a 90% confidence level, the per-bin differences in Figure 4 are significant for the 0.0% WER bin, for the three bins covering the range 15.01% through 30%, and for the bin for 40.01% through 45%. Note that for equal WER, ATC is higher in 2001, which we consider to be an important accomplishment by the research groups who built the Communicator systems.

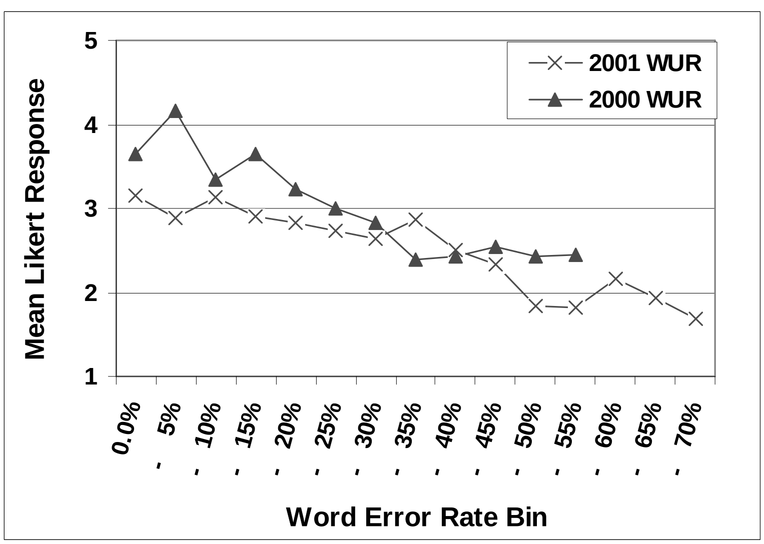

Figure 5. Values of “Would Use Regularly” for 2000 and 2001.

Figure 5 shows that WUR was significantly better in 2000 than in 2001 over the range 0% ≤ WER ≤ 30% ($p < .01$ over that WER range). But the WUR results can be considered equivalent for the two years at higher levels of WER. What we find most interesting in Figure 5 is that in 2001 WUR was almost level up through 35% WER and then fell off. This is discussed further in the section of this paper about correlations of WER with user satisfaction. Figure 5 shows that in 2001, over the range 0% ≤ WER ≤ 35% (containing 80% of the calls that year), WUR was basically independent of WER. The multivariate regression analyses by Walker et al. (2001, 2002) also found that WER was not a particularly influential predictor of user satisfaction.

Figure 6 shows the number of user words and the total elapsed time during the travel planning task. For ease of visual comparison, both metrics have been *linearly* rescaled so that the graph lines for the two metrics overlap maximally. As is apparent, the two metrics are correlated (r-square = 0.57). As can also be seen, efficiency deteriorates as WER becomes higher, which is expected since many ASR errors have to be corrected by the user. Note that both metrics become more variable above a WER of 35%. The corresponding metrics in the data for 2000 (not shown) were also more variable above a WER of 35%. We think this increased variability is due to other factors swamping the effects of WER on task efficiency as WER becomes high.

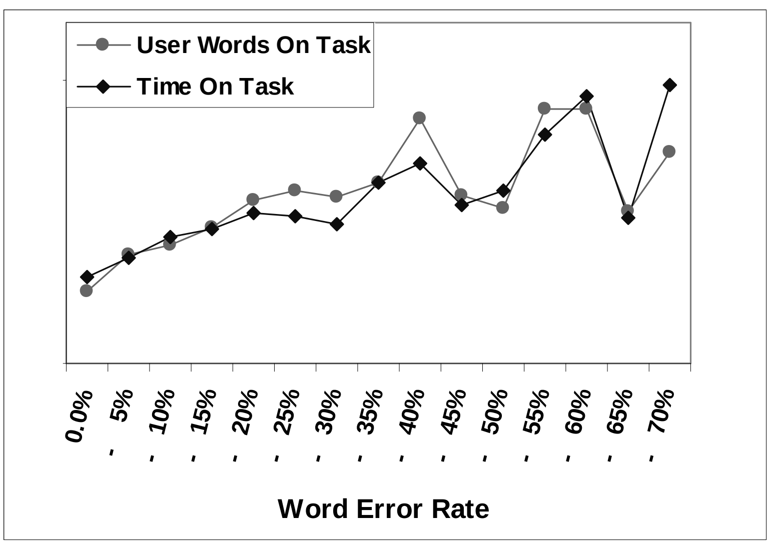

Figure 6. Efficiency metrics by WER histogram bin for the 2001 data.

Correlations of Word Error Rate with Perceived Task Completion

In the 2000 data, there were seven fixed scenario templates, used in order for each subject's first seven calls (template 1 on call 1, etc.). Aggregating the data by scenario template counterbalances for variation across sites and variation across scenario instantiations, making the data less noisy. Further, each template represents a large number of calls and callers.

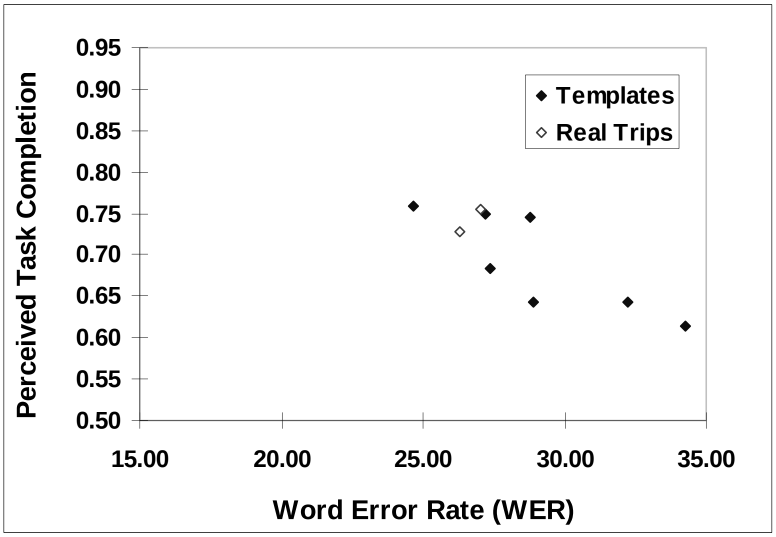

Figure 7. Scatter plot of PTC versus WER for calls in 2000 aggregated by scenario template.

Figure 7 shows PTC vs. WER in 2000, on calls aggregated by scenario template, with the instantiations equally weighted. It includes data points for calls 8 and 9 (the user-created or “real” trips). The closest equivalent for the 2001 data is to aggregate by hypothetical scenario with one added data point representing all the real trips, as shown in Figure 8.

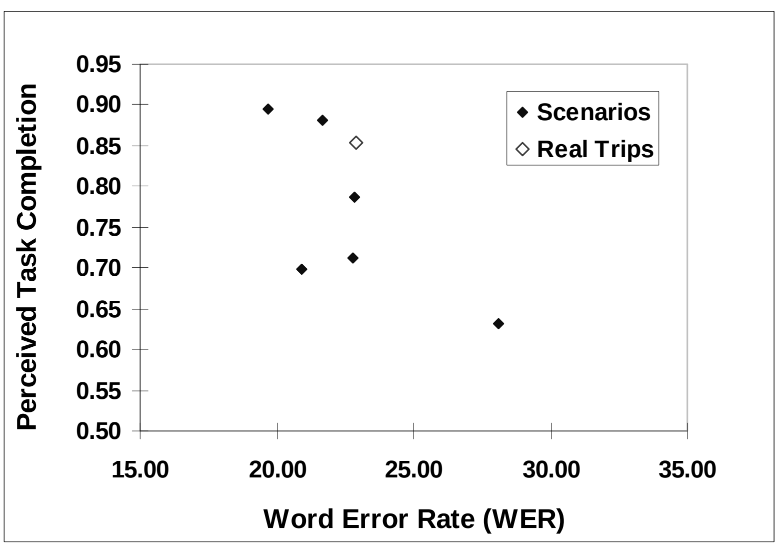

Figure 8. Scatter plot of WER versus PTC for the 2001 hypothetical scenarios.

Although Figures 7 and 8, which show aggregated data, suggest a strong linear correlation, actually least-squares linear-regression over unaggregated data (that is, by call) for the 2000 data, for WER predicting PTC, gives r-square of only 0.14, with slope of only -0.0084 (the slope shows the effect of WER on PTC, so that the effect size is that an increase of 10% in WER would tend to decrease PTC by 8.4%). The 95% confidence interval for the slope is -0.010 through -0.0067. Doing the exactly corresponding regression on the 2001 PTC data gives r-square 0.069, which is quite low, with slope -0.0047 (that is, an increase of 10% in WER would tend to decrease PTC by 4.7%). The 95% confidence interval for the 2001 slope is -0.0057 through -0.0037. Comparing the slopes of these PTC regression models suggests the effect of WER on PTC was only about half as great in 2001 as in 2000.

It is reasonable to ask whether multivariate linear regression would give a better model for PTC over the 2001 data that we have been discussing. We think two of the eleven available metrics, *Mean Words Per System Turn* and *Task Ease*, show the results of task success rather than being causes of it. The *Mean Words Per System Turn* metric is a circular explanation of task completion because the systems normally gave a long readback of the itinerary whenever the task was completed. *Task Ease* ("In this conversation, it was easy to get the information that I wanted") is similar to asking whether the task (getting information) was easy to complete. Nine remaining potential explanatory factors are listed in Table 2.

Expected Behavior,
Interface Confidence,
Num User Utterances,
Num User Words (sum over all user utterances),
Num Overlaps (times when system and user were both speaking),
Sum Of Time In Overlaps,
Time On Task,
WER, and
WUR

Table 2. Potential explanatory factors for Task Completion.

Expected Behavior, *Interface Confidence*, and *WUR*, are Likert-style items from the user questionnaire (Table 1), and the other factors are calculated from the logfiles. We would, of course, like to simplify our model to have fewer than nine predictor variables. Three of these nine variables (*Time On Task*, *Num User Utterances*, and *Num User Words*) tell us much the same thing, and of the three, *Num User Words* is the one that is most strongly correlated with PTC as well as with ATC (in multivariate regression as well as univariate regression). So, of the three, we decided to keep *Num User Words* and eliminate the other two on theoretic grounds as redundant.

In the PARADISE models derived by Walker, et al. (1992), *Time On Task* is an important determinant of User Satisfaction. But *Time On Task* is not an important determinant of PTC: in fact in 2001 there was little difference (9.6 seconds, or about 2%) in mean *Time On Task* for calls with PTC = 0 vs. PTC = 1. And although the difference in mean *Time On Task* over the calls with ATC = 0 vs. ATC = 1 was 87 seconds, or about 20%, if it were included in our multivariate model for ATC it would have about an order of magnitude less influence than *Expected Behavior*. For these reasons, we think eliminating it as redundant with *Num User Words* is not sacrificing useful information.

One could suspect that WER is partially redundant with *WUR*, but users had no direct way to see WER so we think those two are reasonably independent — we will have more to say about the relationship between the two later in the paper. The variables that relate to overlaps, *Num Overlaps* and *Sum Overlaps*, had almost no effect in the model and were eliminated (in fact, just under 10% of the calls had any overlaps at all, so one could argue for eliminating them on principle). This left us with a set of five independent predictor variables. Any of the usual statistical procedures for simplifying the set of predictor variables then further eliminates *Num User Words* and *Interface Confidence* from the set of five factors, as less influential.

Thus, our multivariate linear-regression model for PTC was reduced to three independent variables: *Expected Behavior*, WER, and WUR, resulting in the model given in Table 3 for PTC over the 2001 data, with r-square 0.23 and with 95% confidence interval of 0.44 through 0.56 for the slope of the regression line.

Metric	Coefficient	PTC = 0	PTC = 1	Difference	StdDev	| Coeff.*StdDev |
Exp.Behav.	0.108	1.85	3.48	1.63	1.37	0.148
WER	– 0.0025	34.50	19.94	–14.56	21.30	0.053
WUR	0.013	1.72	2.97	1.26	1.44	0.019

Table 3. Multivariate linear regression model for PTC.

In Table 3, the Metrics and Coefficients columns state the resulting multivariate model:

$$\text{PTC} = (\textit{ExpectedBehavior} * 0.108) - (\textit{WER} * 0.00251) + (\textit{WUR} * 0.0128) + \text{a constant.}$$

The PTC=0 and PTC=1 columns state the mean value of the PTC metric over the unsuccessful and successful calls, respectively. The Difference column is the PTC=1 value minus the PTC=0 value. The StdDev column gives the standard deviations for the values of the metrics over all the calls. The |Coeff*StdDev| column gives some idea of the relative influence of the three independent predictor variables in the model, with *Expected Behavior* turning out to be very much the most influential factor and WUR having perhaps an eighth as much effect. One could consider Coefficient * Difference as an alternative indicator of relative influence. The r-square value (0.23) for the multivariate model in Table 3 is much higher than the r-square value of 0.069 for the univariate model with WER as the sole independent variable. But a value of 0.23 for r-square is, nevertheless, still rather small, indicating that other factors (not captured in our metrics) must be quite influential. To us, the dominance of *Expected Behavior* suggests that the quality of dialogue handling is an obvious candidate for such additional factors. Figure 9 shows the relationship between WER and the other two explanatory factors in this model.

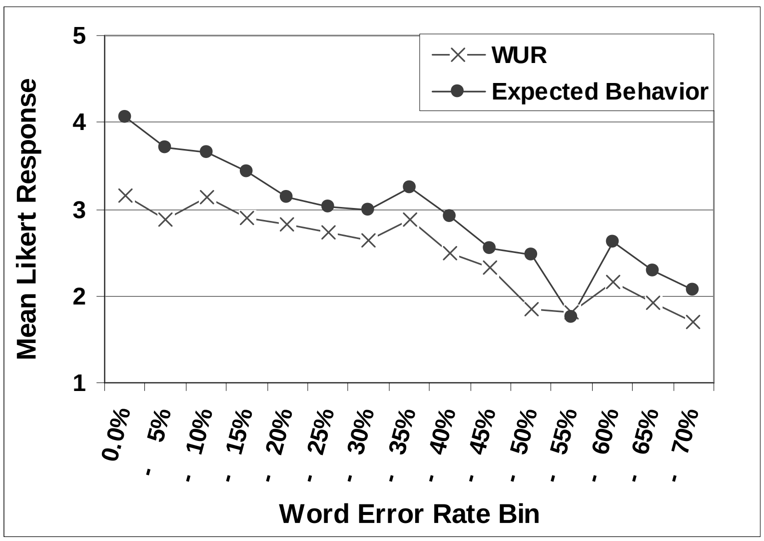

Figure 9. Explanatory factors for Task Completion in the 2001 data (see Table 3 and Table 4).

Correlation of Word Error Rate with Annotated Task Completion

In addition to PTC, we have task completion as annotated by human analysts—by independent annotators at AT&T in 2000 and by the sites and NIST in 2001. We call this metric Annotated Task Completion (ATC). ATC is a more independent objective task completion metric than our subjective task completion metric, PTC. PTC represents the opinion of the user about task completion. ATC, in contrast, represents the opinion of the system developers or of independent annotators. We consider ATC more accurate than PTC because it has stronger correlations with other metrics including user satisfaction, because it has a nicely linear correlation with WER, and especially because it should be comparable (consistent) across all users for a site.

Figures 10 and 11 showing ATC correspond to Figures 7 and 8 showing PTC on the same data. The outlier in Figure 10 represents template one, each user’s first call in 2000, which should have been the easiest template. In contrast, template one in Figure 7 is the point that almost overlaps with the point for a real trip: apparently the users in 2000 overstated their task success rate in PTC on their first call.

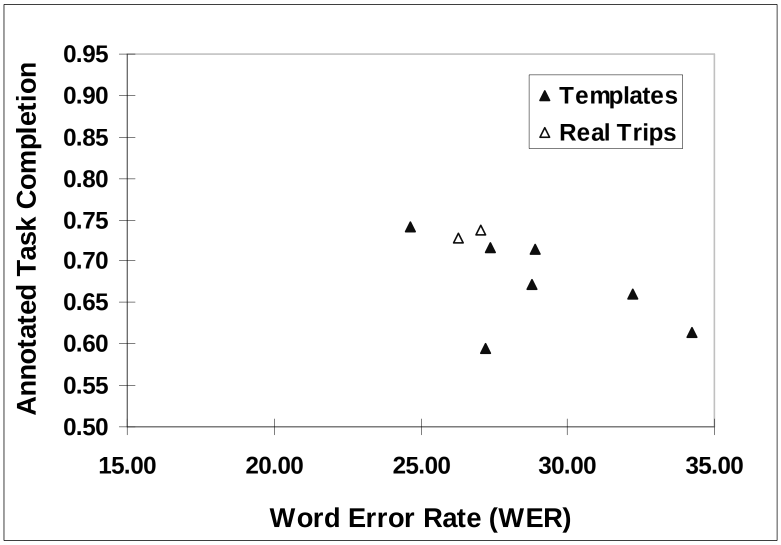

Figure 10. Scatter plot of WER versus ATC on 2000 templates (instantiations equally weighted).

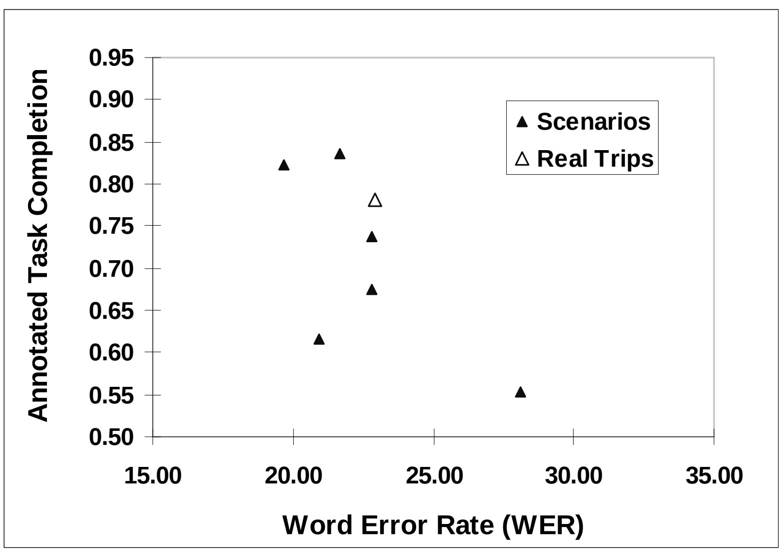

Figure 11. Scatter plot of WER against ATC for 2001 hypothetical scenarios.

On the unaggregated 2000 data, a linear regression of WER with ATC gives r-square 0.12, with the least-squares regression line having slope -0.0091 (10% increase in WER would decrease ATC by about 9.1% in 2000). The 95% confidence interval for the slope is -0.011 to -0.0072. The corresponding linear regression for the 2001 data gives r-square = 0.11 and slope = -0.0067 (10% increase in WER would decrease ATC by about 6.7% in 2001) with 95% confidence interval for the slope of -0.0078 to -0.0056. As the slope for 2000 is not within the 95%

confidence interval for 2001, and vice-versa, we conclude that (like PTC) ATC was significantly less dependent on the quality of ASR in 2001 than in 2000, which we consider to be an important improvement. However, as was mentioned in the discussion of why we did not pool the data from the two years (see Figure 4) the slopes do not differ significantly over the WER range 2% ≤ WER ≤ 55% — the difference in slopes for ATC is mostly a matter of better performance on calls with high WER in 2001.

Figures 7, 8, 10, and 11 show data aggregated by scenario template. Those four figures suggest the templates differ in difficulty for speech recognition. We investigated many hypotheses for why particular templates should give good or bad speech recognition results, but we did not find *any* statistical evidence to support *any* of our explanatory hypotheses. It may be the case that the differences in WER seen among templates in these four figures are random, even if the effects of those WER differences are not random. There are no statistically significant differences in WUR results among the templates.

As was the case for PTC, multivariate linear regression gives a better model for ATC over the 2001 data that we have been discussing. Beginning with the same set of nine independent variables as for PTC, and reducing that set (as previously described when discussing our multivariate regression model for PTC) our model for ATC reduced to the same three independent variables (*Expected Behavior*, WER, and WUR), giving us the following model for ATC on the 2001 data, with r-square 0.23 and with 95% confidence interval of 0.42 through 0.56 for the slope of the regression line.

Metric	Coefficient	ATC = 0	ATC = 1	Difference	StdDev	| Coeff.*StdDev |
Exp.Behav.	0.11	2.18	3.52	1.34	1.37	0.151
WER	– 0.0046	34.74	18.49	–16.25	21.30	0.098
WUR	0.0044	2.02	2.99	0.98	1.44	0.0063

Table 4. Multivariate linear regression model for ATC.

In Table 4, as in Table 3, the Metrics and Coefficients columns state the multivariate model:

$$ATC = (\mathit{ExpectedBehavior} * 0.11 \;-\; (\mathit{WER} * 0.0046) + (\mathit{WUR} * 0.0044) + \text{a constant.}$$

As was the case in our multivariate model for PTC, *Expected Behavior* is very much the most influential factor in our model for ATC. Comparing our multivariate models for ATC and PTC reveals that, in the model for ATC, WER is almost twice as influential, and WUR only about a third as influential, than in the model for PTC. In some sense, this is what one would expect if ATC is a more objective measure of task completion.

As in the multivariate model for PTC, the multivariate model for ATC has a much higher r-square value (0.23) than the r-square value of 0.11 for the univariate model of WER alone predicting ATC. But r-square of 0.23 is, nevertheless, a fairly low value, which suggests (as was the case for PTC) that there are other influential factors not captured in our metrics. The dominance of *Expected Behavior* in the model for ATC (as in the model for PTC) suggests that dialogue handling abilities are an obvious candidate for such factors, as was also the case for PTC.

Correlations of Word Error Rate with User Satisfaction

User satisfaction is an interesting aspect of the systems to measure. The task-completion metrics measure task success, but not whether achieving success was efficient, easy, or pleasant. Nor do they measure how confident the user felt that success would be achieved during the session.

As has been mentioned, we are using WUR as our metric for user satisfaction. Figure 5 showed that WUR values deteriorated when WER exceeded 35% in 2001. We think the correct explanation of this pattern is that other factors swamp the effects of WER as the WER becomes higher and that to some degree the calls with high WER represent different callers than the calls with low WER. We think it possible that the callers who tend to have a high WER may differ from the callers with lower WER in multiple ways, perhaps tending to be more tolerant of ASR difficulties, if their difficulty in being understood by the speech recognition systems corresponds to difficulty in being understood by other people.

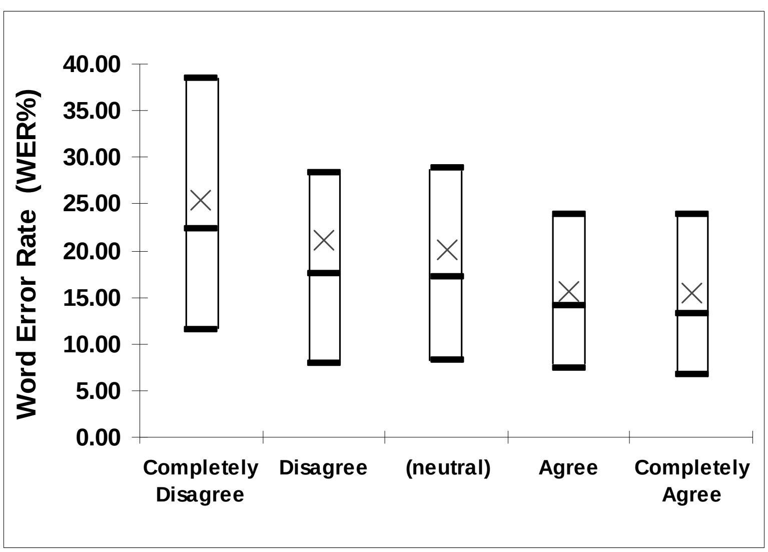

Figure 12. Inter-quartile ranges, medians, and 90% trimmed means of WER by Likert response to WouldUseRegularly (WUR) on 2001 data.

Another way of looking at User Satisfaction is Figure 12, in which the bars show the inter-quartile range of WER for each possible WUR response, with the line across the middle of each bar showing the median, and the X in each bar showing the mid-90% trimmed mean. Poor ASR tended to produce system behavior that led to user dissatisfaction. In Figure 12, the trimmed mean WER for the “Completely Disagree” responses is significantly higher than the trimmed mean WER for the “Disagree” and “Neutral” responses ($p < .01$), which are in turn significantly higher than the trimmed mean WER for the “Agree” and “Completely Agree” responses ($p < .01$). But although the calls with favorable WUR responses (“Agree” and “Completely Agree”) were less likely to have high WER values, one can also see that many calls with unfavorable WUR responses had low WER — good ASR was not sufficient to guarantee high user satisfaction.

Results by site on the same metrics, with WER held constant

As a final look at these data from a different perspective, we return to the question of whether differences among the systems in WER (ASR performance) accounted for their differences in task completion or user satisfaction—which is a hypothesis that was strongly held by some at an earlier stage of the program. If differences in WER are the whole explanation, one would expect that selecting a large subset of calls from the system built by each research site, such that the subsets from the sites have essentially equal means, would give essentially equal results. But as can be seen in Figure 13, that was not the case.

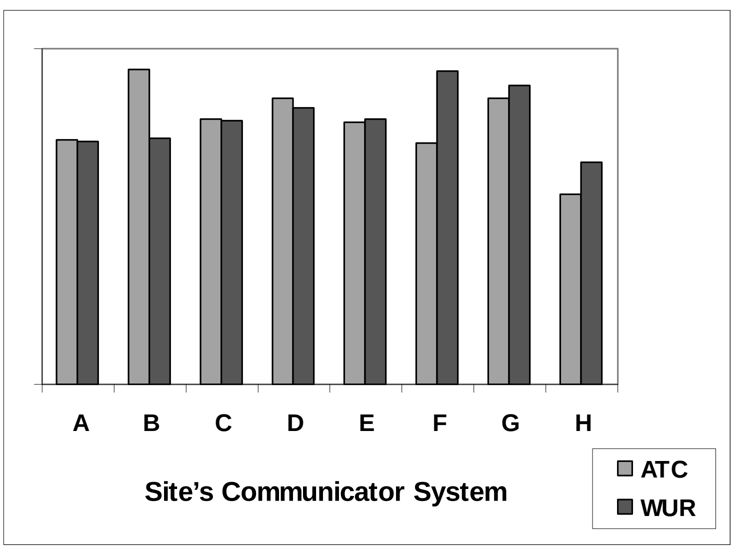

Figure 13. Task Completion and User Satisfaction performance, by site, on large subsets of calls that have equal mean Word Error Rates.

All the sites had many calls in the WER range of 5% through 50%. We selected three overlapping WER bands that cover 5% through 50% WER. For each band we selected the subset of calls in that band from each site, slightly adjusting the WER endpoints of the band for each site's set in order to get essentially equal mean for each site, in each WER band, across the sites. We computed Task Completion (ATC) and User Satisfaction (WUR) for each of the three WER subset and averaged together the three Task Completion and User Satisfaction results to get a composite result for each site. These composite results correspond to a mean WER of about 21% for each site and also represent somewhat similar variances as a result of the "three bands" approach. This particular analysis (perhaps unfairly) deprives sites of the advantage (or disadvantage) of differences in ASR performance in order to show that when one holds WER constant the sites differ in other ways. Figure 13 shows these composite results, with the values rescaled linearly such that the tallest Task Completion bar in the graph is the same height as the tallest User Satisfaction bar. We believe the results shown in Figure 13 are comparable in the sense suggested by the figure caption.

If WER were the entire explanation of either Task Completion or User Satisfaction, we would expect all the Task Completion bars to be about the same height and/or all the User Satisfaction bars to be about the same height. But they are not.

Figure 13 is intended to show patterns, and statistical significance is a bit difficult to assess since the figure shows averages over three overlapping WER bands, which is not an exact match to the assumptions for any of the usual statistical tests. However, reasonable estimates of standard error for these bar heights (using the underlying data for calls with WER of 5% through 50%) suggests there is no statistically significant difference between the ATC bar heights for sites A and F, and none between the ATC bar heights for sites C and E, with the other differences in ATC bar heights appearing significant. Similarly, there is no statistically significant difference between the WUR bar heights for sites A and B, none among the WUR bar heights for sites C, D, and E, and none between the WUR bar heights for sites D and G. All other differences in bar heights in Figure 13 may be statistically significant at the 95% confidence level. Thus, we conclude that there are real differences in both Task Completion and User Satisfaction among the sites that are obviously not accounted for by WER. Factors other than WER have been shown in the multivariate regression for ATC in Table 4 and the analyses by Walker et al. (1991, 1992).

We think additional factors other than WER that account for the differences seen in Figure 13 are successful understanding of the meaning of what the user says, tactics for dealing with users changing course, good models of the user's task domain, and successful tactics for repairs of misunderstandings. We need metrics for these factors (see, e.g., Bernsen, Dybkjær, and Dybkjær, 1997). That is an important topic of future work.

Summary

Accurate speech recognition is a very important factor in task success, and the effect appears linear even to quite high levels of word error rate. Asking users whether they successfully completed their intended task seems to be less accurate than having an independent analyst determine task success.

User satisfaction has important determinants other than the accuracy of speech recognition, and we think the best analysis is given in the papers that have been mentioned by M.A. Walker, et al., using the PARADISE framework. User satisfaction does deteriorate at high levels of WER.

Metrics for the efficiency of performing the task using the system are most interesting in models for user satisfaction. At high values of WER, other factors probably swamp the effects of WER on efficiency. Values of WER above 35% were associated with increased variability of multiple metrics, suggesting that a system's ability to deal with ASR errors becomes more influential above a WER of 35%. But most Communicator calls had relatively low WER — 80% of the calls in 2001 had WER less than 35%, and 95% of the calls had WER less than 65%.

References

Aberdeen, J. (2000). DARPA Communicator Logfile Standard. (http://fofoca.mitre.org/logstandard)

Bernsen, N. O., Dybkjær, H., and Dybkjær, L. (1997). What should your speech system say? *IEEE Computer* 30(12), pp. 25–31.

Bonneau-Maynard, H., Devillers, L., and Rosset, S. (2000). Predictive performance of dialog systems. *LREC 2000, Proceedings of the Second International Conference on Language Resources and Evaluation* (May 31–June 2, 2000). Athens, Greece: European Language Resources Association, pp. 177–181.

Le, A.N., Sanders G.A., and Aberdeen, J. (January 25, 2001). DMA Metrics and Their Log Standard Implementations, Version 5. (Available on request from gregory.sanders@nist.gov or audrey.le@nist.gov).

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9, pp. 19–35.

Polifroni, J., and Seneff, S. (2000). Galaxy-II as an Architecture for Spoken Dialog Evaluation. *LREC 2000, Proceedings of the Second International Conference on Language Resources and Evaluation* (May 31–June 2, 2000). Athens, Greece: European Language Resources Association, pp. 725–730.

Sanders, G. A., and ATIS Committee (March 9, 2001). Communicator Transcription Guidelines, Version 1.2. (Available on email request from gregory.sanders@nist.gov or audrey.le@nist.gov).

Sanders, G.A., Le, A.N., and Garofolo, J.S. (2002). Effects of Word Error Rate in the DARPA Communicator Data During 2000 and 2001. *ICSLP-2002*, 7*th International Conference on Spoken Language Processing* (September 16–20, 2002). Denver, Colorado: International Speech Communication Association, pp. 277–280.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., and Zue, V. (1998). GALAXY-II: A Reference Architecture for Conversational System Development. *ICSLP-1998, International Conference on Spoken Language Processing.* Sydney, Australia: International Speech Communication Association, pp. 931–934.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Walker, M. A., Litman, D.J., Kamm, C.A., and Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the 35th ACL and 8th EACL* (Madrid, Spain). San Francisco, CA: Association for Computational Linguistics and Morgan Kaufmann, pp. 271–280.

Walker, M.A., Kamm, C.A., and Litman, D.J. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 6:363–377.

Walker, M., Aberdeen, J., Boland, J.,Bratt, E., Garofolo, J., Hirschman, L., Le, A.,Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. (2001). DARPA Communicator dialog travel planning systems: The June 2000 data collection. *EUROSPEECH 2001, 7th European Conference on Speech Communication and Technology* (September 3–7, 2001). Aalborg, Denmark: International Speech Communication Association, pp. 1371–1374.

Walker, M., Passonneau, R., Aberdeen, J., Boland, J.,Bratt, E., Garofolo, J., Hirschman, L., Le, A.,Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. (2002). DARPA Communicator Evaluation: Progress from 2000 to

2001. *ICSLP-2002, 7th International Conference on Spoken Language Processing* (September 16–20, 2002). Denver, Colorado: International Speech Communication Association, pp. 273–276.

Walker, M., Passonneau, R., Aberdeen, J., Boland, J.,Bratt, E., Garofolo, J., Hirschman, L., Le, A.,Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. (2002). DARPA Communicator: Cross-System Results for the 2001 Evaluation. *ICSLP-2002, 7th International Conference on Spoken Language Processing* (September 16–20, 2002). Denver, Colorado: International Speech Communication Association, pp. 269–272.