Transcription Conventions and Evaluation Techniques for Spoken Language System Research

Alexander I. Rudnicky Michelle H. Sakamoto

27 November 1989
CMU-CS-89-194 3

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Also circulated as a DARPA Spoken Language Systems Note.

Abstract

We describe the transcription conventions currently in use for spontaneous speech at Carnegie Mellon University. Two sets of conventions are described, a detail-rich system for *wizard* experiments, and a more rigid *evaluation* system designed for purposes of SLS evaluation. The latter is suitable for automatic scoring using the existing NBS (now NIST) scoring software. A sample wizard transcription is included as well as a sample of live-system transcription together with system output. Transcripts can be used to generate a number of diagnostic metrics useful for system evaluation.

The research described in this paper was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, monitored by SPAWAR under contract N00039-85-C-0163. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or of the US Government.

Table of Contents

List of Figures

List of Tables

Begin SLS Note 5

Transcription Conventions and Evaluation Techniques for Spoken Language System Research

Alexander I. Rudnicky and Michelle H. Sakamoto
Carnegie Mellon University School of Computer Science, Pittsburgh, PA 15213
27 November 1989

Abstract

This note describes the transcription conventions currently in use for spontaneous speech at Carnegie Mellon University. Two sets of conventions are described, a detail-rich system for *wizard* experiments, and a more rigid system designed for purposes of SLS evaluation. The latter is suitable for automatic scoring using the existing NBS (now NIST) scoring software. A sample wizard transcription is included as well as a sample of live-system transcription together with system output. Transcripts can be used to generate a number of diagnostic metrics useful for system evaluation.

Spoken language research requires the creation of orthographic versions of recorded speech material, both to facilitate the analysis of spontaneous speech corpora and to allow for system evaluation. This memo describes two sets of conventions for producing orthographic transcriptions of spoken language data. These conventions were developed in the course of research on spoken language systems performed at Carnegie Mellon University and are distinguished by the degree of descriptive detail that each provides. We do not propose any conventions for the phonetic transcription of spontaneous speech at this time, though such would be of value for certain kinds of corpus analysis.

The first set of conventions (the ***wizard*** style) is meant for transcription of speech collected in the course of "wizard" experiments that simulate a spoken language system by means of a human operator, typically hidden from the person using the simulated system. The wizard transcription style allows the researcher to include various annotations, such as commentary and suprasegmental indicators, that go beyond a strict lexically-based transcription but are useful for exploratory analysis.

The second set of conventions (the ***evaluation*** style) is meant for transcription of speech elicited in the course of live interaction with a spoken language system and whose transcription needs to be related to some form of recognition system output. In the evaluation style, many of the feature of the wizard style are unnecessary, since the goal is to compare the transcription with the lexical-level output of a (limited) automatic transcriber. The evaluation style also needs to adhere to a more rigid format which allows for mechanical scoring, such as that provided by existing NBS (NIST) scoring software [Pallett 89].

There is, of course, no reason not to use the latter conventions for the transcription of wizard material, particularly if the transcriptions are then to be used as input to the parsing component of a spoken-language system (say for its evaluation in isolation). It is simply a question of what purpose the transcription is meant to serve.

More generally, it should be understood that it might not be possible to formulate a definitive transcription style, since any one style makes presuppositions about the the use to which it will be put. The best that can be hoped for is that a particular convention will adequately support the needs that it was meant to address and that it is able to comfortably accommodate some range of un-

anticipated uses.

1 Transcription of "Wizard" data

This section describes the conventions in use at Carnegie Mellon for the transcription of speech recorded in the course of "wizard" experiments that explore human-computer interaction by voice. These conventions were developed for the transcription of unconstrained goal-directed speech but would also be suited for more constrained speech.

Our goal in developing these conventions was to provide text data suitable for the following purposes: development of a "speech language" for the implementation of performance tasks (such as a voice spreadsheet), the analysis of spontaneous speech phenomena (such as pauses, restarts, and extraneous events), and for the analysis of prosodic phenomena (such as emphasis and boundary marks). The conventions are derived from a number of sources [Newell and Simon 72, Sacks *et al.* 74] as well as experience gained from non-speech protocol experiments. We believe that they strike the proper balance between detail and abstraction and provide data suitable for a variety of applications.

1.1 The Recording and Transcription Process

We record sessions in an "office" environment, meaning that other activities are taking place while the user performs the task (e.g., other voices, phone ringing, door closing, etc.). The *user* is seated at a monitor (e.g., a Sun console). An *experimenter* interacts with the user, explaining the task and giving directions in case of difficulty. A second person, the *operator*, sits out of sight of the user (either behind the user or across a partition). The task of the operator is to translate spoken commands into appropriate computer commands.

No attempt is made (in these particular experiments) to mislead the user about the supposed use of a recognition system, since we are interested in obtaining data under unconstrained conditions, where the user feels free to chose the most natural form of expression. It is of course possible to contrive a situation in which the user is lead to believe that he or she is interacting with an actual recognition system, as in e.g., [Hauptmann and Rudnicky 88], and to produce a rather different style of interaction. The choice depends on the goal being pursued.

We record speech using a Nikko D-100III cassette recorder. Some of the sessions were recorded using a Realistic PZM microphone (Radio Shack) placed next to the computer terminal. The intent was to leave the participant as unencumbered as possible. We found that this produced recordings of sufficiently high quality for transcription (that is, no portions of the tapes were unintelligible). In a second study, we switched to a close-talking microphone (Sennheiser HMD-224), with the intent of being able to digitize the recordings for further analysis. Transcription was done using a Dictaphone 2870 transcription machine. A machine built specifically for transcription greatly simplifies the task and is highly recommended. The material is typed directly into the computer, using a text editor. To catch and correct the inevitable errors, a second person listens to the tape and verifies the correctness of the transcription.

1.2 Description of Codes

The following speech and session event codes are used. The coding scheme was chosen to allow manual analysis as well as some forms of automatic processing (for example, as input to a parser capable of handling spontaneous speech phenomena).

`<cr>` Line breaks delimit single utterances that (often) correspond to complete com-

mands. The text could just as easily be considered as a single stream. However, line breaks impose a meaningful segmentation on the material and thereby increase the readability of the transcript for humans.

`[*]` Indicates the point in time at which the operator typed in a command. Typically, though not invariably, these occur at line breaks. This symbol might be thought to represent "system response". If more detailed information is needed (e.g., the system response itself), it can be included (e.g., `[* "a system response"]`).

Colon (`:`) Indicates lengthening, typically of a vowel sound. The colon is usually placed immediately after the sound that is lengthened

Hyphen (`tw-`) A word *ending* with a hyphen indicates that the speaker cut that word short. If the identity of the word is not obvious from the fragment, then the transcription may specify the intended word (e.g., `tw[enty]-`) if such is obvious to the transcribers. Word-internal hyphens have no special significance. Interrupted but continued words are coded with a hyphen following the first part of the word (e.g., `hy- phen`).

Period (`.`) Indicates silence, each period corresponding roughly to one second of elapsed time. Note that only pauses internal to an utterance (line) are coded.

Comma (`,`) Indicates a boundary mark, either a short pause or an inflection. The placement of a comma reflects in part a subjective judgment.

Capitalization (e.g., `CLOthing`)
Indicates emphatic stress. That is, stress beyond what might normally be expected on the basis of lexical or syntactic factors.

Square brackets (e.g., `[rustle]`)
Describes extraneous audible events that are sufficiently loud to potentially impact a recognition system. Sessions can include some amount of interaction with the experimenter, for example, when a problem comes up ("uh, my screen just disappeared"). Experimenter comments are either transcribed verbatim, or the interaction is summarized (and placed in brackets). Brackets also enclose general information ("comments") that might be included in a transcript.

1.3 A full Wizard transcript

The transcript in Table 4 is included to demonstrate the use of the transcription system for a simulated spreadsheet data entry task. The data are for one particular speaker (`a.h.`). The total elapsed time for this session is about 11 minutes. Note that the time includes the substantial pauses that correspond to the interval during which the operator types in a command. Based on other experiments, we estimate that this session would have taken about half the time to complete, had "system response" been instantaneous.

2 Transcription conventions for system evaluation

The transcription style described in the previous section provides a rich description of spontaneous speech. The style described in the present section is meant to facilitate mechanical evaluation of recognizer output. The conventions presented below are currently being used at Carnegie Mellon for the evaluation of a live spoken language system (for a description, see [Rudnicky, et al. 89]). The categories we have developed are those that we have found to be of use in trying to understand our system and to work on improvements. Different categorizations are possible, both more broad and more detailed. The appropriate level of detail depends, of course, on the uses to

which the data will be put.

Compared to the evaluation of recognition systems developed under the just-concluded DARPA speech recognition program (see, e.g., [Pallett 89]), the evaluation of spoken language recognition is problematic for two reasons: First, the lexical items encountered will not be part of a closed set that can be specified *a priori*. Second, it is not possible to create a definitive reference for each utterance.

Such problems do not typically arise when systems are developed using read speech data, since it is possible to create a completely specified correspondence between the symbols generated by the recognizer and the symbols used to (exhaustively) describe the contents of the utterances.[1] In the case of a spoken language system being evaluated in live situations this is no longer true, since various acoustic events (whether speech or non-speech) which are not explicitly modeled by the system may occur as input. Trivially, the problem could be dealt with by assigning all such events to some cover symbol (e.g., ++UNKNOWN+). However to do so would lose much of the diagnostic information that could be of use in understanding system performance. We therefore believe that some attempt should be made to classify these events.

Spontaneous speech also presents the problem of determining exactly what was spoken in a particular utterance. For read speech, the intended utterance is specified in advance and depending on the care with which the recording sessions are conducted, utterances that do not seem to instantiate the reference can be either re-recorded or can be eliminated from the corpus. No such reference exists for live speech, since the "intention" for a given utterance is generated on the fly by the system user. For most utterances this is not a problem, though cases of ambiguity do exist. An example might be the distinction between "HUNDRED AND NINETY" and "HUNDRED NINETY", where the presence of a reduced AND may be difficult to ascertain. In such cases, we have to rely on the judgment of the transcribers and on the explicitness of transcription guidelines.

2.1 The transcription process

We define an *event* as audible acoustic energy delimited by silences or by other labeled events. When two events overlap, preference is given to the lexically meaningful element (e.g., word over noise), or to the element attributable to the nominal session talker. Otherwise, the most salient event (as judged by the transcriber) is given preference. No attempt is made to further code overlapping events. We place events in live speech into one of three categories: **lexical, extra-lexical,** and **non-lexical**. These will be explained in greater detail below.

To provide consistency in the transcription process, the following guidelines were developed:

- *Transcribe all words*. If a particular word is not recognizable, a guess is made, based on the transcriber's best understanding of the context of occurrence, both sentential and task, in which the word occurs. If a word or phrase cannot be identified with reasonable confidence, then the "++MUMBLE+" marker may be used. If a word is mispronounced but is nevertheless recognized correctly, it is transcribed as if it were spoken correctly. If it is misrecognized, it is transcribed as heard.
- *Label all other audible events*. At the least level of detail, these can be identified by a cover symbol ("++NOISE+"). We have found, however, that it is useful to label separately those events that occur frequently enough to be of interest in themselves,

---

[1] In previous work, the correspondence has not been strictly one-to-one at the symbol level. The evaluation system therefore included well-defined rules for mapping non-standard items into the recognizer reference set of symbols.

such as breath noises, or perhaps telephone rings. Table 1 lists those symbols we have introduced for labeling the live Spreadsheet Corpus.

- *If the system recognizes an interrupted word correctly, then the word is transcribed as if it were spoken in its entirety.* This convention is arbitrary and obviously hides information about interrupted words that are nevertheless correctly recognized. It may need to be revised as we begin to explicitly study such phenomena.
- *Utterances that produce no recognizer output are eliminated from the transcript.* In our system, this consisted of zero-length "utterances" that result from malfunctions in endpoint detection. A record of these utterances should of course be kept, so that relevant statistics can be calculated.
- By convention, *any utterances at the beginning of the session that reflect the user's unawareness that the system just went live are eliminated.* This speech consists of interactions with the experimenter and typically reflect the user's unawareness that the period of instruction has ended. Since the user is not "using" the system, this convention is justified. A record of such deletions is kept, however. We have also encountered one case in which the user kept interacting with the system after the end of a task, intentionally "testing" the system with out-of-task material. These utterances were eliminated from the session in question.
- *Extraneous noises are always transcribed if they affect the recognition.* Otherwise, noises are transcribed only if, in the opinion of the transcriber, they are sufficiently prominent ("loud enough"). Certain noises, in particular inhalations and exhalations at the start and end of an utterance, are not typically transcribed. To determine the validity of this convention, we informally examined the output of a recognizer trained to detect extraneous events [Ward 89]. We found no inconsistencies.

The transcription of live-session material was done using a NeXT workstation. All our speech data were kept on NeXT "floptical" disks, each of which could hold about two complete user datasets (about 1500 utterances each). The transcriber listened to the speech using "walkman" type open earphones. A simple utility was written to present the transcriber with the utterance, the corresponding recognizer output and a copy of the latter (in an editor buffer) which was to be corrected to correspond to the recorded speech. Since word accuracy was generally quite high (ranging from 79.8% to 94.8% and averaging 90.1%), the editing procedure (as opposed to blind transcription) resulted in substantial time savings. Note of course that this produces some degree of bias in favor of the system, since its output is used as a template. In our judgment this bias is not significant for purposes of evaluation. To verify the correctness of the transcription, a second person listened to all utterances, comparing them against the transcription. The checker did not have access to the system response during verification.

Table 1: Non-lexical items coded in the Spreadsheet Corpus

1.3326	585	++RUSTLE+	
0.4692	206	++BREATH+	
0.0980	43	++MUMBLE+	
0.0410	18	++SNIFF+	
0.0296	13	++BACKGROUND-NOISE+	
0.0251	11	++MOUTH-NOISE+	
0.0228	10	++COUGH+	
0.0137	6	++YAWN+	
0.0114	5	++GIGGLE+	
			0.0091 4 ++PHONE-RING+
			0.0091 4 ++NOISE+
			0.0091 4 ++DOOR-SLAM+
			0.0091 4 ++CLEARING-THROAT+
			0.0091 4 ++BACKGROUND-VOICES+
			0.0046 2 ++SNEEZE+
			0.0023 1 ++SIGH+
			0.0023 1 ++PING+
			0.0023 1 ++BACKGROUND-LAUGH+

Note: The first column gives the percent of all tokens represented by the listed item. The second column gives the actual count, the third column lists the transcription item.

2.2 Description of codes

The following conventions were designed to meet a number of goals. We needed an accurate rendition of what was said. We also needed a format that would simply mechanical processing. Finally, the coding scheme needed to be compatible with the existing NBS scoring software. We use the following notational conventions:

HUNDRED	Words in the lexicon are transcribed using exactly the form defined by the system lexicon. This is known as a **lexical** item.
+THANKS+	An out of vocabulary word is bracketed by + symbols. No distinction is made between words directed at the system and words directed at humans present in the environment (such as the experimenter). Clearly, such a distinction can be made, if necessary (e.g., by using the +++ marker described below). This is referred to as an **extra-lexical** item.
+TW-	An interrupted word is bracketed by a + and a -, to indicate that it is a fragment of a larger, lexically appropriate item. By convention, this is known as an **extra-lexical** item.
++RUSTLE+	An extraneous event descriptor begins with ++ and ends with a +. We have differentiated descriptors for frequently occurring events (such as paper rustles). Low frequency events could be described by more general labels, such as ++NOISE+. This is referred to as a **non-lexical** item.
+++GRAMMAR+	Any additional annotations, such as a marker indicating the parsability of a particular utterance, use a prefix of +++. These markers are to be ignored in any analysis. If additional marker categories are needed (and the idea of indefinitely long strings of +s does not appeal), the prefix can be of the form +*n*+, where *n* is a number (using letters would create ambiguity).

2.3 Characteristics of the transcribed corpus

For the initial portion of the Spreadsheet Corpus (consisting of 15 voice sessions from 8 different talkers), a total of 12,507 utterances were transcribed, containing 43,901 lexical tokens. There were 212 unique tokens. Table 3 gives the distribution of these tokens over the three categories described above. Perhaps surprisingly, over half of the items (57%) fall outside the lexicon, though these constitute only about 2.5% of tokens transcribed.

To assess the accuracy of our transcription procedure, one complete session transcript was reviewed in detail by a panel consisting of the transcriber, the checker, and two others. Disagreements were found for the labeling of extraneous events, but these differences were deemed to be of marginal importance. Since only two individuals were involved in the transcription process (the transcriber and the checker), we believe that the extraneous-event labels, if perhaps not completely consistent with the intuitions of others, are certainly internally consistent.

To provide a quantitative assessment of transcription accuracy, the authors (one of whom was the checker) listened to a further five sessions, comparing the transcription with the recorded speech. This validation set contained of a total of 699 utterances and 2360 words. We found only two errors: One word was omitted from a long digit string and one non-lexical item was not transcribed. The latter (a click) should probably have been transcribed, since it produced an insertion error. We therefore estimate that the error rate for word-level transcription is about 0.1%. The utterance transcription error would be about 0.3%[2]. We are satisfied that this level of transcription

[2]Correcting the two transcription errors noted above there would not, however, have changed the utterance error rate for the set we examined, since both utterances in question were errorful for other reasons.

Table 2: Extra-lexical items coded in the Spreadsheet Corpus

0.0342 15 +GO+	0.0023 1 +WHOO-	0.0023 1 +MY+
0.0273 12 +AH+	0.0023 1 +WHATS+	0.0023 1 +MEDI-
0.0182 8 +DOLLARS+	0.0023 1 +WHAT+	0.0023 1 +LA-
0.0114 5 +S-	0.0023 1 +UH+	0.0023 1 +KNOW+
0.0091 4 +TW-	0.0023 1 +U-	0.0023 1 +JUST+
0.0091 4 +SA-	0.0023 1 +TRILLION+	0.0023 1 +JESUS+
0.0091 4 +FUCK+	0.0023 1 +THO-	0.0023 1 +IT+
0.0091 4 +FI-	0.0023 1 +THE+	0.0023 1 +INVESTMENT+
0.0068 3 +THIS+	0.0023 1 +THATS+	0.0023 1 +INS-
0.0068 3 +SHIT+	0.0023 1 +THANK+	0.0023 1 +IN+
0.0068 3 +P-	0.0023 1 +T-	0.0023 1 +IM+
0.0068 3 +LIVING-EXPENSES+	0.0023 1 +T+	0.0023 1 +HOWS-THAT+
0.0068 3 +DAMMIT+	0.0023 1 +SUNK+	0.0023 1 +HOW+
0.0046 2 +YOU+	0.0023 1 +STO-	0.0023 1 +HERE+
0.0046 2 +THING+	0.0023 1 +SON-OF-A+	0.0023 1 +HAH+
0.0046 2 +TH-	0.0023 1 +SHIP+	0.0023 1 +H-
0.0046 2 +SEV-	0.0023 1 +SEVEN-	0.0023 1 +GRRR+
0.0046 2 +SE-	0.0023 1 +SETS+	0.0023 1 +GOING+
0.0046 2 +N-	0.0023 1 +SALES+	0.0023 1 +FIFT-
0.0046 2 +MED+	0.0023 1 +SAL-	0.0023 1 +FI--VE+
0.0046 2 +LAB+	0.0023 1 +SAINT+	0.0023 1 +EMERGENCY+
0.0046 2 +I+	0.0023 1 +REALIZED+	0.0023 1 +DRED+
0.0046 2 +HUND-	0.0023 1 +PL-	0.0023 1 +D-
0.0046 2 +GO-	0.0023 1 +PHEW+	0.0023 1 +CREACT+
0.0046 2 +G-	0.0023 1 +PERSONAL-PROPER-	0.0023 1 +CAR-INSURANCE+
0.0046 2 +FUCKING+	0.0023 1 +OY+	0.0023 1 +BATTLE+
0.0046 2 +FIRE+	0.0023 1 +OUTA+	0.0023 1 +AW-SHOOT+
0.0046 2 +DICK+	0.0023 1 +OOPS+	0.0023 1 +AW-FUCK+
0.0046 2 +DAMN-YOU+	0.0023 1 +OO+	0.0023 1 +AUTOMOBILE+
0.0046 2 +A-	0.0023 1 +ON+	0.0023 1 +ASS-
0.0023 1 +Z-	0.0023 1 +OH-OK+	0.0023 1 +AHH+
0.0023 1 +WRONG+	0.0023 1 +OH-NO+	0.0023 1 +ACCOUNT+
0.0023 1 +WOW+	0.0023 1 +OH+	0.0023 1 +ABOUT+
0.0023 1 +WITH+	0.0023 1 +OF+	0.0023 1 +A+

Note: The first column gives the percent of all tokens represented by the listed item. The second column gives the actual count, the third column lists the transcription item.

accuracy is adequate for system evaluation, given our current absolute error levels of recognition error and the inter-user variance in this error rate (see section).

Table 3: Lexical characteristics of the Spreadsheet Corpus

Item type	type count	type incidence	token count	token incidence
Lexical	92	43%	42,802	97.5%
Extra-lexical	102	48%	177	0.4%
Non-lexical	18	9%	922	2.1%

2.4 A full Live transcript

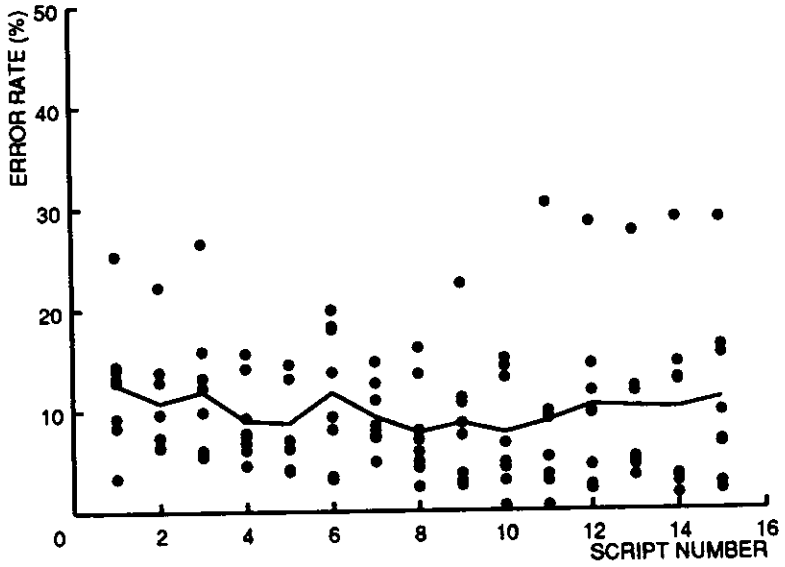
Table 5 displays the complete transcript of a session completed by speaker `cps` doing spreadsheet task 18. All utterances, defined as an activation of the recognizer, are numbered sequentially (`cps-de18.`*n*). The gaps in the numbering sequence correspond to "empty" recognitions, occasioned, for example, by a failure in the end-point detection algorithm.

3 The mechanics of evaluation

3.1 SLS output with corresponding transcription

Table 5 shows an extract from the output produced by the NBS scoring program [Pallett 89]. The REF lines show the transcription while the HYP lines show the recognizer output. The run shown in this Table is one that would be used for diagnostic purposes. For example, note that transcribed non-lexical items cause the scoring algorithm to produce an error, even if the recognizer correctly transcribed the intended utterance. The summary statistics generated for such a run can be used to detect some patterns in the data, such as which words are typically matched to a given non-lexical item. (We noticed, for example, that paper rustles were often recognized as the word THREE.)

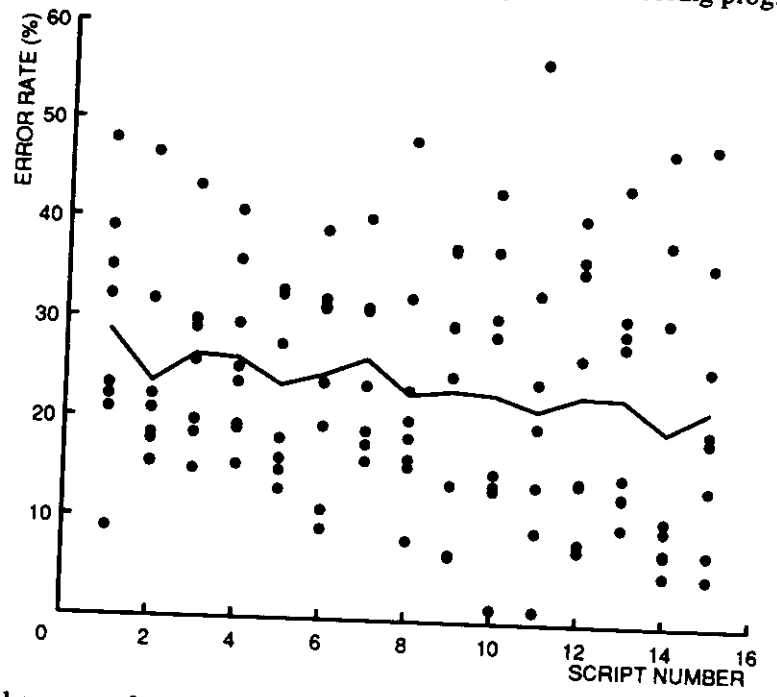
Figure 1: Word Accuracy calculated using the NBS scoring program



The amount of data generated by spoken language systems in use for tasks lends itself to the calculation of meaningful performance statistics. Figures 1 and 2 show values for two of the standard metrics provided by the NBS program, word accuracy and sentences correct. The plotted points correspond to a single session. Each is therefore based on about 100 utterances. The line corresponds to the mean error rate. Over all 15 sessions, the mean word accuracy is 90.1% (with a standard deviation of 6.6) and the mean sentences correct is 76.2% (σ 11.9). As can bee seen, word accuracy does not appear to vary across sessions: recognizer performance does not improve at the word level with continued use. Sentence accuracy seems to improve over sessions. This trend, however, is not statistically significant (as determined through an analysis of variance). Differences between users, on the other hand, are significant and account for about half of the variance in the sample. We have no reason to believe that the high variances exhibited by these data are in any way unusual. As such, they suggest caution in interpreting differences in performance between different systems or even between versions of the same system.

The focus of evaluation for a spoken language system should be on how the system performs as a whole and how efficiently it allows the user to perform a given task. The above statistics, although useful for understanding the performance of the speech recognition component of an SLS are not adequate for characterizing system response. In the next section, we propose several metrics that

Figure 2: Sentences Correct calculated using the NBS scoring program



quantify additional aspects of system performance.

3.2 Scoring the SLS output

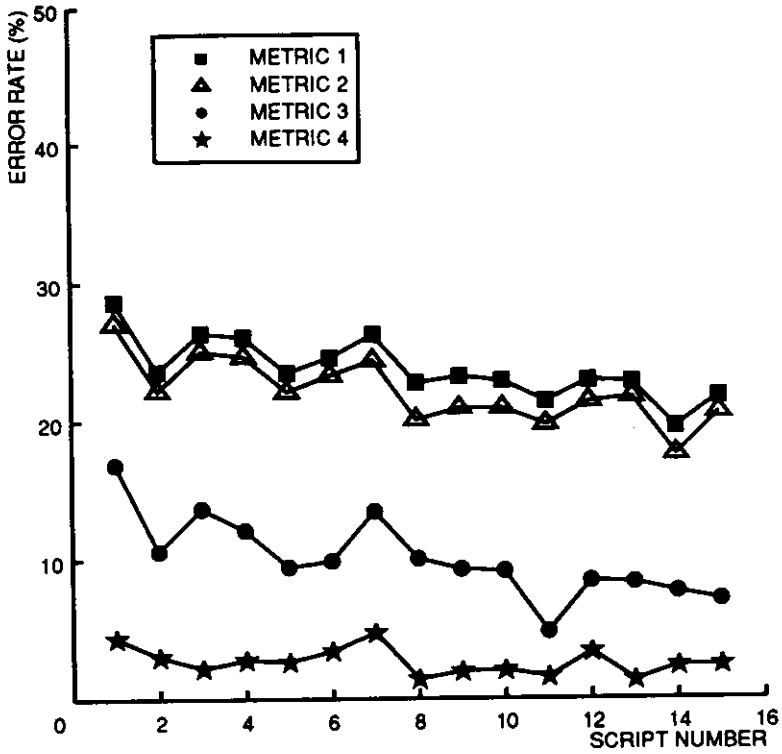
The transcription conventions defined in this document allow us to calculate a variety of statistics that characterize the performance of a spoken language system, based on transformations of the reference transcript. Arguments for the appropriateness of these metrics are presented elsewhere (e.g., [Rudnicky, et al. 89]). The purpose of the present discussion is to describe how these metrics are calculated. Figure 3 shows system performance (in terms of utterance error) for eight users, over a series of 15 spreadsheet tasks.

The different curves are defined as follows:

Metric 1 — **Exact performance** is calculated by using the NBS alignment program to compare the manual transcription with the corresponding string produced by the recognizer. Note that this criterion produces a very conservative estimate of system performance. A more realistic estimate of spoken language system performance is provided by the second metric.

Metric 2 — The **semantic error** rate is calculated by determining whether the (low level) goal in force at the time of the utterance was successfully achieved. The spreadsheet task that we have studied is well-specified in terms of a goal structure. Users enter a succession of items, each entry requires a positioning action followed by an insertion action, etc. Because of this it is possible to accurately determine the success of each interaction, since failure is apparent from repeated attempts at satisfying the current goal. Utterances were marked accordingly. As can be seen from the Figure, the current system provides little recovery from recognition error, not surprisingly so, since there is no global semantic component in the system *per se* (the improvement shown derives

Figure 3: Outcomes of different recognition scoring procedures



from constraints imposed by the word-pair grammar used in the recognizer). More sophisticated systems that attempt to reinterpret recognitions, say in terms of their understanding of the user's intentions, would exhibit (we believe) a substantial spread between the exact and semantic curves. Other systems, such as ones that apply (semantic) constraints in the course of the recognition itself might not show this spread, unless these constraint mechanisms are disabled. We further believe that this spread between exact and semantic accuracy represents a useful quantification of the additional power provided by the higher-level components (such as semantic and pragmatic) of an SLS, and can serve as a useful metric for tracking SLS performance at this level.

Metric 3

The **extraneous event** rate is calculated from the transcription and show the percentage of utterances that contain material not strictly interpretable by the system parser. Such failure is caused by the presence of either extra-lexical or non-lexical items. This particular curve allows us to determine the "cleanliness" of the speech in a corpus, incorporating a measure of both how well users manage to stay within the language specified by the system (for both grammar and lexicon), and how well they manage to control the occurrence of non-lexical items. Note that the current data indicate that users progressively learn to control their input to the system, halving the number of corrupted utterances by the end of the measuring period.

Metric 4

The last metric indicates the **grammatical error** rate. This is calculated by eliminating all non-lexical (++) items from the transcription and determining whether the remaining strings pass through the system parser. Grammaticality

(or *coverage*) indicates the number of utterances that lie outside the language, assuming that the system can handle all non-lexical items by some other means. The current spreadsheet system provides coverage of about 97% over the course of the 15 sessions. This is quite high and very likely reflects the inherent constraints imposed by the rather simple task that users were asked to perform. A somewhat different pattern might have been observed if the task involved higher-level communication with the system, for example a planning task for which the system was expected to implement the consequences of an abstractly specified constraint.

Note that each of the above curves can be easily generated by simple filtering operations over the transcriptions. The information for the semantic accuracy must be produced manually at the time the original transcription is created. By removing various classes of + tokens from the reference transcriptions, the analyses for Metrics 3 and 4 can be easily performed.

4 Summary

This note has described two transcription styles suitable for spoken language research, together with a sample evaluation. Transcription styles are arbitrary, their content being governed by the needs dictated by ongoing research. We have found that the current styles meet our needs. We have also described some summary statistics for a corpus of spreadsheet data, including lexical characteristics and recognizer performance. We also discussed four separate evaluation metrics and demonstrated their usefulness for understanding system performance and language characteristics.

5 Acknowledgments

A number of people have contributed to the work described in this paper. We would like to thank Joseph Polifroni who collected and transcribed the bulk of the wizard data and Takima Hoy who transcribed the bulk of the live session data.

6 References

[Hauptmann and Rudnicky 88]
Hauptmann, A. G. and Rudnicky, A. I. Talking to computers: An empirical investigation. *International Journal of Man-Machine Studies* 28:583-604, 1988.

[Newell and Simon 72]
Newell, A. and Simon, H.A. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[Pallett 89] Pallett, D.S. Benchmark tests for DARPA Resource management database performance evaluations. In *Proceedings of ICASSP*, pages 536-539. IEEE, May, 1989.

[Rudnicky, et al. 89]
Rudnicky, A.I., Sakamoto, M.H., and Polifroni, J.H. Evaluation spoken language interaction. In *Proceedings of the DARPA Workshop on Spoken Language Systems*. October, 1989.

[Sacks *et al.* 74] Sacks, H., Schegloff, E. A., and Jefferson, G. A Simplest Semantics for the Organization of Turn-Taking for Conversation. *Language* 50(4):696-735, 1974.

[Ward 89] Ward, W.H. Modelling Noise Events with HMMs. In *Proceedings of the DARPA workshop on spoken language systems*. October, 1989.

7 Transcripts

Transcripts of sample Wizard and Live sessions, and a sample NBS alignment follow.

Table 4: Transcript of Wizard session for spreadsheet task

[a. h.; person 1]
oka:y
u:h go to the cell for my salary [*]
and enter the amount, six thousand five hundred [*]
'kay go to the cell for my RENT, in the income section [*]
and enter the amount five fifty [*]
u:h . go to the STOCKS, line of the, dividends, section [*]
and enter the amount one hundred fifty-eight dollars and fifty cents [*]
u:h .. okay . go down to the savings, line in the interest section [*]
and enter the amount fifty-four [*]
okay .. u:h .. go down the next screen [*]
okay, go to mortgage payment ..
a:nd enter the amount seven hundred forty-eight dollars and fifty-seven cents [*]
u:h .. okay, on the next, line
enter the amount, two hundred forty three dollars and twenty-seven cents [*]
okay . under BANK charges .. u:h enter the amount fifteen dollars [*]
and under credit card enter the amount zero [*]
okay .. on the electricity line under utilities [*]
enter the amount . nineteen dollars and forty-seven cents [*]
on the next line, enter six tw[enty]- sixty-two twenty-five [*]
on the next line, enter seventy-five ninety-four [*]
on the next line, enter six twenty [*]
a:nd on the next line, enter fifteen [*]
hm ... [sigh]
okay ... uh, go down another screen [*]
okay . ADD a line betwee:n, the entertainment line and the food line . under living expenses [*]
okay .. a:nd label this item, movies, and indent it, so it's a subsection of entertainment [*]
okay . now for that . section enter the amount, uh forty dollars [*]
okay
uh . could you . refresh the screen [*]
okay . u:h .. under restaurant,
enter two hundred seventeen eighty-five [*]
oh . oops .. change that to h- two forty-six [*]
[breath] and now, uh add another subsection between lines forty-eight and forty-nine [*]
labeled .. hosting [*]
and .. f- for that subsh- subsection, fill in the amount two hundred seventeen eighty-five [*]
[breath] .. okay
uh under clothing .. enter ..... [the] sum .. o:f . six hundred 'n fifty .. fifty ... twenty-five ... that's it. [*]
u:h .... okay, under GAsoline .. enter, sixty-five [*]
o:kay
under child support and day care .. on line thirty:,eight i guess it is ... enter . two thousand five hundred [*]
okay [*]
uh go down a screen [*]
screen's worth [*]
okay, under, oh, under house MAINtenance [*]
um .. enter the amount a hundred ninety-five [*]
under, car maintenance, enter fifteen ninety-five [*]
[click] u:h
okay
go up one screen [*]
okay under the category medical, enter forty-five [*]
oka:y, now go down a screen again [*]
okay, under car insurance, enter seventy dollars [*]
under house insurance enter fifty-two dollars [*]

u:h under .. miscellaneous dues [*]
enter seventy-five [*]
okay, find the taxes section [*]
show that
okay
under federal, ente:r . four seven two three point nine one [*]
under state enter, three seventy-eight point thirty-four [*]
a:nd under city, enter twenty-three forty-five [*]
okay, find my ASSets [*]
o:kay, can i see a little more of that [sigh] [*]
okay
u:m
okay under house,
enter a hundred sixty-seven thousand, even [*]
under automobiles, enter seventeen thousand five hundred even [*]
please refresh the screen [*]
[exhalation] hm
[exhalation]
okay
ADD eighty-five thousand even, to houses [*]
u:h .... o:kay
under, personal property .. enter the sum of five hundred e:ven and twelve hundred e:ven [*]
under checking enter nine forty-five point six seven [*]
u:nder debts receivable [*]
enter zero [*]
under sa- ving [throat clearing] ente:r, two thousand eight hundred and seventy-six, and twenty-five cents [*]
oh ... woops . i'm sorry change that to:,
the sum .. of two eight seven six point two fi:ve [*]
four thousand five hundred e:ven and two thousand dollars even [*]
u:h hm:
i'm not sure i did that correctly
[oper: change it]
yeah . change that back to twenty-eight seventy-six point two five [*]
okay, no:w ADD another ROW between eighty-six and eighty-seven [*]
ca:lled, um stocks and bonds [*]
and, for that amount, enter the sum o- of forty-five hundred and two thousand [*]
oh .. woops [laugh] okay delete that entire row
[laugh]
okay, go down another screen [*]
okay, under STOCKS, enter forty-five hundred [*]
under bonds, enter two thousand [*]
unde:r retirement account, enter ten thousand [*]
hm
okay
go down to . another screen's worth .. please [*]
okay, under MORTgage,
ente:r, a hundred twenty-five thousand [*]
under car loan, enter nine thousand seven hundred [*]
unde:r . credit card balance [*]
enter four four two seventy-three [*]
point seventy-three i'm sorry [*]
and under .... charge account balance, enter two seventy-six point twenty-three [*]
[inhalation]
okay
[says doesn't know how to do 25\% discount]

Table 5: Transcription of Live session for spreadsheet task

cps-del8.2	++breath+ goto salary
cps-del8.3	goto salary
cps-del8.4	goto salary
cps-del8.5	goto salary
cps-del8.6	goto salary
cps-del8.7	goto b six
cps-del8.8	goto b six
cps-del8.9	goto b six
cps-del8.11	seven thousand eight hundred
cps-del8.13	seven thousand eight hundred
cps-del8.14	goto rent
cps-del8.15	five hundred and seventy ++rustle+
cps-del8.16	goto stocks
cps-del8.17	four hundred and one point one
cps-del8.18	goto savings
cps-del8.19	twelve point o seven
cps-del8.20	twelve point o seven
cps-del8.23	twelve point zero seven
cps-del8.24	goto mortgage-payments
cps-del8.25	six hundred forty one point three three
cps-del8.26	six hundred forty one point three three
cps-del8.27	goto car-payments
cps-del8.28	hundred forty two point four seven
cps-del8.30	one four two point four seven
cps-del8.31	goto bank-charges
cps-del8.32	++breath+ ten ++sniff+
cps-del8.34	ten
cps-del8.35	goto electricity
cps-del8.36	thirty four point eight four
cps-del8.37	down
cps-del8.38	thirty nine ++rustle+
cps-del8.39	down
cps-del8.40	ninety three point six one
cps-del8.41	down
cps-del8.43	six point two five
cps-del8.44	down
cps-del8.45	twelve point nine five
cps-del8.46	goto entertainment
cps-del8.47	fifty six point four five
cps-del8.48	goto restaurant
cps-del8.49	one six three point eight seven five plus five seven point three eight
cps-del8.50	++rustle+ goto gasoline
cps-del8.52	seventy five
cps-del8.53	goto clothing
cps-del8.54	two times fifty plus three times twenty five
cps-del8.55	goto child-support
cps-del8.56	one thousand two hundred
cps-del8.57	goto b ninety eight
cps-del8.58	goto b fifty eight
cps-del8.59	goto b fifty eight
cps-del8.60	left two
cps-del8.61	left two
cps-del8.63	back two
cps-del8.64	hundred fifty eight
cps-del8.65	down
cps-del8.66	nineteen ninety five
cps-del8.68	nineteen ninety five
cps-del8.69	one nine point nine five
cps-del8.70	goto medical
cps-del8.73	thirty two point five
cps-del8.74	goto +car-insurance+
cps-del8.75	down two
cps-del8.76	fifty two point one
cps-del8.77	up
cps-del8.78	twenty nine point three nine
cps-del8.79	goto contributions

```
cps-del8.80     fifty
cps-del8.81     goto dues
cps-del8.82     forty two
cps-del8.83     goto federal
cps-del8.85     one nine one three point seven two
cps-del8.86     down ++breath+
cps-del8.87     down
cps-del8.88     goto state
cps-del8.89     two three four point four five
cps-del8.90     up two
cps-del8.91     goto city
cps-del8.92     goto city
cps-del8.93     thirty seven  point zero one
cps-del8.94     thirty seven point zero one
cps-del8.95     goto automobiles
cps-del8.96     six zero seven five
cps-del8.97     six zero seven five
cps-del8.98     six thousand seventy five
cps-del8.99     goto  houses
cps-del8.101    goto houses
cps-del8.102    eighty three thousand nine hundred
cps-del8.103    eighty three thousand nine hundred
cps-del8.104    goto
cps-del8.105    eighty three thousand nine hundred plus seventy five thousand
cps-del8.106    ++rustle+ goto personal-property
cps-del8.107    ten thousand two hundred
cps-del8.108    ++rustle+
cps-del8.111    down
cps-del8.112    one one one nine point eight two
cps-del8.113    up two
cps-del8.114    seven eight three point nine
cps-del8.115    seven eight three point nine
cps-del8.117    seven eight three point nine
cps-del8.118    seven eight three point nine
cps-del8.119    down ten
cps-del8.120    goto b ninety nine
cps-del8.121    three four zero five
cps-del8.122    down
cps-del8.124    down
cps-del8.125    one five six
cps-del8.126    goto retirement-accounts
cps-del8.127    seven thousand five hundred
cps-del8.128    seven thousand five hundred
cps-del8.130    seven thousand five hundred
cps-del8.131    seven thousand five hundred
cps-del8.132    goto mortgage
cps-del8.133    fifty eight thousand
cps-del8.134    goto car-loan
cps-del8.135    five hundred and sixty nine point eight eight
cps-del8.137    five hundred sixty nine point eight eight
cps-del8.138    goto credit-card-balance
cps-del8.139    three seven three point o eight plus one one three point six one
cps-del8.140    +oh+ standby
cps-del8.143    standby
```

Table 6: Live session transcript for spreadsheet task, with recognitions.

cps-del8.2	REF: ++BREATH+ GOTO SALARY HYP: HUNDRED POWER EIGHT
cps-del8.3	REF: goto SALARY **** HYP: goto CELL RENT
cps-del8.4	REF: GOTO SALARY HYP: **** LEFT-STRING
cps-del8.5	REF: goto SALARY **** HYP: goto CELL RENT
cps-del8.6	REF: goto SALARY **** HYP: goto CELL RENT
cps-del8.7	REF: goto B SIX HYP: goto * GIFTS
cps-del8.8	REF: GOTO B SIX HYP: **** * ***
cps-del8.9	REF: goto b six HYP: goto b six
cps-del8.11	REF: **** seven thousand EIGHT hundred HYP: FIVE seven thousand ***** hundred
cps-del8.13	REF: seven thousand eight hundred HYP: seven thousand eight hundred
cps-del8.14	REF: goto rent HYP: goto rent
cps-del8.15	REF: five hundred AND seventy ++RUSTLE+ HYP: five hundred *** seventy *********
cps-del8.16	REF: goto stocks HYP: goto stocks
cps-del8.17	REF: four hundred AND one point one HYP: four hundred *** one point one
cps-del8.18	REF: goto savings HYP: goto savings
cps-del8.19	REF: **** twelve point o seven HYP: BACK twelve point o seven
cps-del8.20	REF: ** twelve point o seven ***** HYP: UP twelve point o seven THREE
cps-del8.23	REF: twelve point zero seven HYP: twelve point zero seven
cps-del8.24	REF: goto mortgage-payments HYP: goto mortgage-payments
cps-del8.25	REF: **** six HUNDRED forty one point three three HYP: FOUR six OVER forty one point three three
cps-del8.26	REF: six hundred forty one point three three HYP: six hundred forty one point three three
cps-del8.27	REF: goto car-payments HYP: goto car-payments
cps-del8.28	REF: HUNDRED forty two point four seven HYP: R forty two point four seven
cps-del8.30	REF: one four two point four seven HYP: one four two point four seven
cps-del8.31	REF: goto bank-charges

```
                HYP: goto bank-charges

cps-del8.32     REF: ++BREATH+ ten ++SNIFF+
                HYP: LEFT      ten THOUSAND

cps-del8.34     REF: ten
                HYP: ten

cps-del8.35     REF: goto electricity
                HYP: goto electricity

cps-del8.36     REF: thirty four point eight four
                HYP: thirty four point eight four

cps-del8.37     REF: down
                HYP: down

cps-del8.38     REF: thirty nine ++RUSTLE+
                HYP: thirty nine *********

cps-del8.39     REF: down
                HYP: down

cps-del8.40     REF: ninety three point six one
                HYP: ninety three point six one

cps-del8.41     REF: down
                HYP: down

cps-del8.43     REF: six point two five
                HYP: six point two five

cps-del8.44     REF: down
                HYP: down

cps-del8.45     REF: twelve point nine five
                HYP: twelve point nine five

cps-del8.46     REF: goto entertainment
                HYP: goto entertainment

cps-del8.47     REF: fifty six point four five
                HYP: fifty six point four five

cps-del8.48     REF: goto restaurant
                HYP: goto restaurant

cps-del8.49     REF: one six three point eight seven five plus five seven point three eight
                HYP: one six three point eight seven five plus five seven point three eight

cps-del8.50     REF: ++RUSTLE+ goto gasoline
                HYP: ********* goto gasoline

cps-del8.52     REF: seventy five
                HYP: seventy five

cps-del8.53     REF: goto clothing
                HYP: goto clothing

cps-del8.54     REF: two times fifty plus three times twenty five
                HYP: two times fifty plus three times twenty five

cps-del8.55     REF: goto child-support
                HYP: goto child-support

cps-del8.56     REF: one thousand two hundred
                HYP: one thousand two hundred

cps-del8.57     REF: goto b ninety eight
                HYP: goto b ninety eight

cps-del8.58     REF: goto b fifty EIGHT
                HYP: goto b fifty *****

cps-del8.59     REF: goto B fifty eight
                HYP: goto D fifty eight

cps-del8.60     REF: LEFT        TWO
                HYP: LEFT-STRING ***
```

```
cps-del8.61     REF: LEFT          TWO
                HYP: LEFT-STRING ***

cps-del8.63     REF: back two
                HYP: back two

cps-del8.64     REF: hundred fifty eight
                HYP: hundred fifty eight

cps-del8.65     REF: down
                HYP: down

cps-del8.66     REF: NINETEEN * ninety five
                HYP: LET      G ninety five

cps-del8.68     REF: *** NINETEEN ninety five
                HYP: LET T        ninety five

cps-del8.69     REF: one nine point nine five
                HYP: one nine point nine five

cps-del8.70     REF: goto medical
                HYP: goto medical

cps-del8.73     REF: thirty two point five
                HYP: thirty two point five

cps-del8.74     REF: goto +CAR-INSURANCE+
                HYP: goto INSURANCE

cps-del8.75     REF: down two
                HYP: down two

cps-del8.76     REF: fifty two point one
                HYP: fifty two point one

cps-del8.77     REF: up
                HYP: up

cps-del8.78     REF: twenty nine point three nine
                HYP: twenty nine point three nine

cps-del8.79     REF: goto contributions
                HYP: goto contributions

cps-del8.80     REF: fifty
                HYP: fifty

cps-del8.81     REF: goto dues
                HYP: goto dues

cps-del8.82     REF: forty two
                HYP: forty two

cps-del8.83     REF: goto federal
                HYP: goto federal

cps-del8.85     REF: one nine one three point seven two
                HYP: one nine one three point seven two

cps-del8.86     REF: DOWN ++BREATH+
                HYP: F    *********

cps-del8.87     REF: ** DOWN
                HYP: UP ZERO

cps-del8.88     REF: goto state
                HYP: goto state

cps-del8.89     REF: two three four point four five
                HYP: two three four point four five

cps-del8.90     REF: up TWO
                HYP: up TEN

cps-del8.91     REF: GOTO CITY
                HYP: **** EXIT

cps-del8.92     REF: goto city
```

```
                HYP: goto city

cps-de18.93     REF: THIRTY SEVEN  POINT ZERO ONE
                HYP: ****** TWELVE ***** **** ***

cps-de18.94     REF: thirty seven point zero one
                HYP: thirty seven point zero one

cps-de18.95     REF: goto automobiles
                HYP: goto automobiles

cps-de18.96     REF: SIX ZERO SEVEN FIVE
                HYP: *** **** ***** ****

cps-de18.97     REF: SIX ZERO SEVEN FIVE
                HYP: *** **** ***** ****

cps-de18.98     REF: six thousand seventy five
                HYP: six thousand seventy five

cps-de18.99     REF: **** GOTO ***** HOUSES
                HYP: FOUR TWO  POWER S

cps-de18.101    REF: goto houses
                HYP: goto houses

cps-de18.102    REF: ******* EIGHTY three thousand NINE HUNDRED
                HYP: ARC-TAN E      three thousand **** ADD

cps-de18.103    REF: eighty three thousand nine hundred
                HYP: eighty three thousand nine hundred

cps-de18.104    REF: goto ***
                HYP: goto END

cps-de18.105    REF: eighty three thousand nine hundred plus seventy five thousand
                HYP: eighty three thousand nine hundred plus seventy five thousand

cps-de18.106    REF: ++RUSTLE+ goto personal-property
                HYP: ********* goto personal-property

cps-de18.107    REF: ten thousand two hundred
                HYP: ten thousand two hundred

cps-de18.108    REF: ++RUSTLE+
                HYP: QUIT

cps-de18.111    REF: down
                HYP: down

cps-de18.112    REF: one one one nine point eight two
                HYP: one one one nine point eight two

cps-de18.113    REF: up two
                HYP: up two

cps-de18.114    REF: SEVEN   EIGHT three point nine
                HYP: SEVENTY ***** three point nine

cps-de18.115    REF: ***** seven eight three point nine
                HYP: FIFTY seven eight three point nine

cps-de18.117    REF: SEVEN   EIGHT three point NINE
                HYP: SEVENTY ***** three point ONE

cps-de18.118    REF: seven eight three point nine
                HYP: seven eight three point nine

cps-de18.119    REF: down ten
                HYP: down ten

cps-de18.120    REF: goto b ninety nine
                HYP: goto b ninety nine

cps-de18.121    REF: three four zero five
                HYP: three four zero five

cps-de18.122    REF: DOWN
                HYP: F
```

```
cps-del8.124    REF: down
                HYP: down

cps-del8.125    REF: one five six
                HYP: one five six

cps-del8.126    REF: goto retirement-accounts
                HYP: goto retirement-accounts

cps-del8.127    REF: **** seven thousand five hundred
                HYP: FOUR seven thousand five hundred

cps-del8.128    REF: seven thousand five HUNDRED
                HYP: seven thousand five MILLION

cps-del8.130    REF: seven thousand FIVE hundred
                HYP: seven thousand FOUR hundred

cps-del8.131    REF: seven thousand five hundred
                HYP: seven thousand five hundred

cps-del8.132    REF: goto mortgage
                HYP: goto mortgage

cps-del8.133    REF: fifty eight thousand
                HYP: fifty eight thousand

cps-del8.134    REF: goto car-loan
                HYP: goto car-loan

cps-del8.135    REF: five HUNDRED AND SIXTY NINE POINT EIGHT EIGHT
                HYP: five ******* ADD ***** **** ***** ***** *****

cps-del8.137    REF: five hundred sixty nine point eight eight
                HYP: five hundred sixty nine point eight eight

cps-del8.138    REF: goto credit-card-balance
                HYP: goto credit-card-balance

cps-del8.139    REF: three seven three point o eight plus one one three point six one
                HYP: three seven three point o eight plus one one three point six one

cps-del8.140    REF: +OH+ STANDBY
                HYP: FOUR ADD

cps-del8.143    REF: standby
                HYP: standby
```

End SLS note 5