Discovering Cues to Error Detection in Speech Recognition Output: A User-Centered Approach

LINA ZHOU, YONGMEI SHI, DONGSONG ZHANG, AND ANDREW SEARS

LINA ZHOU is an Assistant Professor in the Department of Information Systems at the University of Maryland, Baltimore County. She received a Ph.D. in Computer Science from Peking University, Beijing, China. Her current research interests center on ontology learning, deception detection, natural language processing, and online community. Her work has been published in the *Journal of Management Information Systems, Communications of the ACM, IEEE Transactions on Speech and Audio Processing, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Professional Communication, Decision Support Systems, Information & Management, Group Decision and Negotiation,* and others.

YONGMEI SHI is a Ph.D. Candidate in Computer Science at the University of Maryland, Baltimore County, where she received an M.S. in Computer Science. Her research interests include natural language processing, speech recognition, information retrieval, and machine learning.

DONGSONG ZHANG is an Assistant Professor in the Department of Information Systems at the University of Maryland, Baltimore County. He received a Ph.D. in Management Information Systems from the University of Arizona. His current research focuses on mobile computing, multimedia-based e-learning, computer-mediated communication, and intelligent systems. His work has been published in the *Communications of the ACM, IEEE Transactions on Multimedia, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Professional Communication, Decision Support Systems, Information & Management, Information Systems Frontier, Communications of the AIS, Journal of the American Society for Information Science and Technology,* and others.

ANDREW SEARS is a Professor of Information Systems and the Chair of the Information Systems Department at University of Maryland, Baltimore County. He received his B.S. in Computer Science from Rensselaer Polytechnic Institute, Troy, New York, in 1988 and his Ph.D. in Computer Science from the University of Maryland, College Park, in 1993. He is a coeditor of *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (Mahwah, NJ: Lawrence Erlbaum, 2003). His research explores issues related to human–computer interaction with recent projects investigating issues associated with mobile computing, speech recognition, IT accessibility, and the difficulties IT users experience as a result of their work environment or tasks. Dr. Sears is currently serving as the Adjunct Chair for Education for the Special Interest Group on Computer–Human Interaction and the Treasurer of the Special Interest Group on Accessible Computing.

Journal of Management Information Systems / Spring 2006, Vol. 22, No. 4, pp. 237–270.

0742–1222 / 2006 $9.50 + 0.00.

ABSTRACT: The great potential of speech recognition systems in freeing users' hands while interacting with computers has inspired a variety of promising applications. However, given the performance of the state-of-the-art speech recognition technology today, widespread acceptance of speech recognition technology would not be realistic without designing and developing new approaches to detecting and correcting recognition errors effectively. In seeking solutions to the above problem, identifying cues to error detection (CERD) is central. Our survey of the extant literature on the detection and correction of speech recognition errors reveals that the system-initiated, data-driven approach is dominant, but that heuristics from human users have been largely overlooked. This may have hindered the advance of speech technology. In this research, we propose a user-centered approach to discovering CERD. User studies are carried out to implement the approach. Content analysis of the collected verbal protocols lends itself to a taxonomy of CERD. The CERD discovered in this study can improve our knowledge on CERD by not only validating CERD from a user's perspective but also suggesting promising new CERD for detecting speech recognition errors. Moreover, the analysis of CERD in relation to error types and other CERD provides new insights into the context where specific CERD are effective. The findings of this study can be used to not only improve speech recognition output but also to provide context-aware support for error detection. This will help break the barrier for mainstream adoption of speech technology in a variety of information systems and applications.

KEY WORDS AND PHRASES: cues to error detection, speech recognition, taxonomy, verbal protocol analysis.

SPEECH RECOGNITION IS ONE OF THE MAIN information technologies that provide users with hands-free interaction with computers. The power of speech control promises to help many individuals who might otherwise not be able to interact with a computer through the conventional mouse or keyboard interface. Particularly, users who are physically challenged, visually impaired, or mobility impaired are offered new opportunities by speech recognition–enabled applications.

Although many interesting applications have emerged or enhanced with the advance of speech technology, the current use of such a technology reflects "only a tip of the iceberg of the full power that the technology could potentially offer" (cf. [8, p. 69]). However, some fundamental and practical limitations of the technology result in unsatisfactory performance of speech recognition systems. The convenience and efficiency promised by speech technology in interacting with computers is seriously compromised by the laborious efforts and frustration experienced in detecting and correcting recognition errors [40]. Widespread acceptance of speech as a primary input modality for computers will not be possible unless the underlying recognition technology can produce sufficiently robust and low-error output [8].

To bridge the gap between what people expect from speech recognition and what the technology can achieve, it is desirable to find effective ways to detect and even

correct recognition errors. Error detection is the premise for error correction. No matter whether it is systems or users who are responsible for error detection, they rely on cues to error detection (CERD). In the system-initiated approach, CERD are extracted automatically from internal parameters or the output of speech recognition, and then used to build machine learning models using the training data [5, 47, 50]. The machine learning models are then used to detect possible recognition errors and even to improve speech recognition output (e.g., [21, 47, 49]). Thus, this approach is characterized as data-driven because CERD are selected primarily based on their impact on the training data and their automation potentials. This type of approach is efficient; however, its effectiveness may be undermined by overlooking other significant CERD based on user experience and heuristics knowledge, which are neither automatically generated nor predefined.

According to our survey of the related literature, the data-driven approach is dominant in identifying CERD. The ultimate goal of speech recognition is to approximate native speakers' recognition capability. Given the unsatisfactory performance of systems-initiated error detection and the impact of human factors on information technology implementation [23], there is a strong need to learn CERD from users in order to improve the usefulness of speech recognition systems. However, the extant studies have failed to consider users as an important source for CERD. Therefore, the research question that we aim to address in this study is: What types of CERD do users apply in detecting speech recognition errors?

We propose a knowledge-driven, user-centered approach to discovering CERD in this paper. Instead of relying on trial-and-error, as shown in the systems-initiated, data-driven approach, the proposed method uncovers important CERD from users' knowledge and experience. To the best of our knowledge, this is the first study to discover CERD via a user-centered approach. An empirical study was designed and conducted to elicit verbal explanations from participants about strategies and cues used to detect speech recognition errors, which were then interpreted and encoded via verbal protocol analysis. This type of analysis lends itself to a taxonomy of CERD, consisting of linguistic, hypotheses-based, and other information. Moreover, the association between CERD and three types of speech recognition errors—insertion, deletion, and substitution errors—was analyzed to gain insights into the effectiveness of CERD for specific types of errors. Furthermore, the result of correlation analysis between different CERD suggests that CERD strongly associated with one another can be used together to enhance the performance of error detection.

The developed CERD taxonomy and related findings of this research can make multifold contributions to the detection and correction of speech recognition errors. First, the taxonomy of CERD is the first of its kind. It helps advance our knowledge on CERD and allows future expansion and refinement. Second, it complements CERD discovered by the traditional data-driven approach. Third, the analysis of CERD, in relation to error types, allows us to contextualize CERD to maximize their utility. Fourth, the findings enable the development of supportive and context-sensitive environments to facilitate users in their detection of errors. They enhance the performance of error detection systems with additional knowledge.

Background and Related Work

SPEECH RECOGNITION TECHNOLOGY HAS ADVANCED significantly and has had a tremendous impact on individuals and businesses in the past decade. However, a fundamental challenge that has been taking center stage is the detection and correction of recognition errors. Error detection aims to predict whether there is any mismatch between the output of a speech recognition system and the corresponding reference. Effective error detection relies on knowledge support, which is referred to as cues to error detection.

Benefits and Challenges of the Speech Recognition Technology

Speech technology affects business and human life in ways that go far beyond offering an alternative input mode. It can serve as a tool to support content-based retrieval of audio and audio/visual data [35] and information extraction [13]. Business benefits derived from speech recognition include significantly lowered operational costs; decreased "dropped" calls; improved business agility, customer satisfaction, and loyalty; and increased productivity [8, 19, 30]. Customer services supported by speech recognition technology allow customers' calls to be directed quickly and seamlessly. In addition, the speech recognition technology is especially attractive to the following groups of users:

- People with physical or situation-induced (e.g., on-the-move) impairment. Those who suffer from workplace-associated repetitive stress maladies [2], carpal tunnel syndrome, spinal cord injuries [40], and other similar problems can greatly benefit from this technology by improving their accessibility to computers through the use of their voices rather than a keyboard/mouse. People who are deaf or hard of hearing are enabled to process information presented orally [3]. Moreover, senior citizens, who experience reduction in their mobility, can access computers and applications at the rate of their speaking rather than typing.
- Students with learning disabilities. It is challenging for students with learning disabilities to keep pace with other students in regular classroom settings. Several research projects have incorporated the speech recognition technology into classrooms [3, 18, 25]. The spoken lecture could be simultaneously processed by a speech recognizer, which can serve as an alternative to traditional notetaking. Speech recognition has been proven to have a remedial effect on letter recognition, word recognition, spelling, reading comprehension, phonological processing, and writing fluency [18, 25], which may otherwise be very laborious and frustrating.
- Doctors, lawyers, transcriptionists, and office personnel. Specialized speech recognition software enables more efficient generation and maintenance of the paperwork required in medical and legal fields [22].
- Users of mobile handheld devices. Interaction with handheld devices through a small physical or soft keyboard is slow and error prone. Moreover, handheld devices are often used while traveling, in which typing may be difficult due to

influences such as shaking [16]. Multimodal mobile applications combine voice and touch as input in order to enhance users' experience [48].

- Operators of transportation systems. By preventing attention from being diverted by button or touch screen interaction with hands-free control, speech recognition technologies can improve the safety of vehicle drivers.

Despite the numerous potential benefits of speech recognition technology and continuous research efforts, the accuracy of current speech recognition systems is still far inferior to humans' recognition performance. Speech recognition is challenged by co-occurring ambient noise, continuous utterance, diversity in individual speakers' pronunciation, out-of-vocabulary words, and so on. This has seriously hindered the wide adoption of the technology and its impact on society. No user would like to use a speech recognizer that does not guarantee acceptable levels of recognition accuracy. Given the above challenges in advancing fundamental speech technology, we identified two promising directions for improving the usefulness and adoption of this technology: (1) empowering a speech recognition system with the ability to automatically detect and correct errors, and (2) developing an information system to support users in error detection and correction, which would otherwise be a cumbersome and time-consuming process. Both directions can be pursued by advancing the state-of-the-knowledge on CERD.

Related Work

Minimizing recognition errors remains as a long-standing goal in speech recognition research. CERD are essential to any solutions for detecting recognition errors, which is shown in the extant work related to CERD (e.g., [5, 34, 47, 50]). Moreover, our survey of related studies reveals that the system-initiated, data-driven paradigm is widely adopted.

Confidence Measures

Confidence measures are referred to as a method for detecting hypothesized words that are likely to be erroneous by estimating word and sentence correctness [14]. The result of a confidence measure is denoted by confidence scores. They enable a speech recognition system and subsequent modules to spot the positions of possible errors in the system output automatically [46].

In order to design a confidence measure, one should consider four factors—the level of confidence measure, error definition, predictors used and combination mechanism, and evaluation [6]. A confidence measure can be computed at different levels such as phone [6], word [6, 26], concept [33], and sentence/utterance [32]. It commonly takes into account the values of an array of predictors or CERD. CERD can be extracted from original models in a recognizer or from additional models, which are then combined with either probabilistic or nonprobabilistic mechanism [46]. In order to compute posterior probabilities in the probabilistic approach, word lattice, a compact representation of alternative hypotheses [26], and *n*-best list of top hypotheses

have been used [43]. The majority of nonprobabilistic methods for confidence measures use selected features or CERD to build classifiers for predicting the correctness of a hypothesis. Therefore, CERD are a key component of confidence measures.

Cues to Error Detection

A variety of CERD have been exploited and incorporated into confidence measures in the data-driven approach. According to their generality, CERD can be classified into two categories—recognizer independent and recognizer dependent [50]. Recognizer-independent CERD are generic to different types of speech recognizers. For example, based on the speech recognition output, part-of-speech information [41] can be generated via linguistic analysis. Other types of linguistic information that have served as CERD include syllable, content words [41], parsing mode (i.e., whether a word can be parsed by the grammar and the specific slot position of a parsed word) [49], probability of nodes/arcs assigned in a parse tree, scores from semantic structured language models [37], discourse information [41], and so on. Moreover, linguistic information in the local context has been employed to detect errors [9, 37]. In Sarikaya et al. [37], for example, various context lengths were factored into the computation of semantic structured language model scores. Another group of generic cues come from intermediate parameters generated by a speech recognizer such as acoustic model scores [36, 49], language model scores [36, 44], and posterior word probabilities [20, 44, 46]. Acoustic models are developed to represent audio signals and language models are used to predict the probability of a word based on previous words. Posterior word probability is found to be the best single feature among the internal parameters of a speech recognizer [46]. Moreover, confidence scores of words in the immediate neighborhood influence the probability of the current word being an error [17].

The availability of recognizer-dependent CERD is contingent upon specific speech recognizers. For example, speech recognition output can be presented in several alternative formats—best hypothesis, *n*-best list, and word lattice. Alternative hypotheses (e.g., quickly, quietly, quirkily) are available in an *n*-best list and a word lattice but not in the best hypothesis. If an output word appears rarely in top-*n* alternative utterance hypotheses, that word is likely to be an error. Thus, path ratios (the ratio between number of paths containing a word and the total number of paths) [7, 15, 36] are also employed as CERD.

The above work contributes to our understanding of possible CERD as well as our design of an empirical study for discovering CERD from users.

A User-Centered Approach to Discovering CERD

Despite the notable effect of CERD on improving speech recognition output, as discussed in the previous section, error detection and correction remains as a bottleneck in improving the productivity of speech recognition technology. Two avenues for advancing error detection or correction are identified through our survey of related work. One is to advance the body of knowledge on CERD by learning from human

users. The other is to improve the underlying approaches to confidence measures. A better understanding of CERD is conducive to the effectiveness of confidence measures. Therefore, we focus on the first issue in this research.

As discussed in the introduction, the user-centered approach can potentially overcome the limitations of the systems-initiated, data-driven approach. Although some studies involved human users in error detection, the focal interest of those studies was to either improve the usability of speech recognition systems by building multimodal interfaces [42] or compare user performance in error detection in different settings [41]. Tacit CERD that users apply in error detection and correction remain underexplored. The above situations are accounted for by several factors: (1) it is labor intensive to elicit knowledge from users, (2) the rapid advance in powerful computer technologies has cultivated the tendency to ignore users' roles, and (3) there is a gap between what each research community (e.g., speech technology and human–computer interaction) wants and what it can deliver.

Previous work that bears close relevance to this research was done by Brill et al. [4], who aimed to improve the state-of-the-art language modeling by incorporating more sophisticated linguistic and world knowledge from people. Several major limitations of the study were: (1) the scope of CERD was constrained to linguistic knowledge only; (2) a list of possible CERD was precompiled for participants to choose from, which was essentially a CERD validation process rather than an elicitation process, possibly limiting the type of information that users apply in identifying errors; and (3) the CERD obtained were at a coarse granularity (e.g., argument structure), which are difficult to directly apply in practice.

To advance our knowledge of CERD, we designed a user-centered approach and executed it in an empirical study. In a knowledge-driven, user-centered approach, CERD are elicited from users when they detect and interpret speech recognition errors in real time. Supplementary data (e.g., alternative hypotheses) were also provided to increase the chance of discovering useful knowledge. Based on the related work on CERD, we selected the following types of data to support error detection in this study:

- Speech recognition output is a rich source of CERD. It enables the application of linguistic knowledge and contextual information in error detection as in database applications [10].
- Alternative hypotheses and associated information serve as additional references for judging the correctness of a recognized word.
- In light of the merit of discourse information in error detection, background scenario information about the original speech is likely to be useful in error detection. Nonetheless, existing discourse CERD (e.g., previous dialog act [41]) are designed specifically for spoken dialogue rather than monologue applications such as dictation. New discourse CERD are required to support error detection in dictation.

We did not include parameters of internal models of a speech recognizer to support user error detection because (1) they have already been incorporated while generating

the recognition output, (2) they are rarely accessible in a commercial recognition system, and (3) they are difficult for nonexpert users to use. This study aims to not only test the effectiveness of extant CERD for user error detection but also discover new CERD that can improve error detection by both users and systems.

Research Methodology

WE CONDUCTED LABORATORY EXPERIMENTS to discover CERD that users apply in detecting speech recognition errors. In order to collect data about participants' decisions and reasoning, the think-aloud protocol [11] was employed during the experiments to let participants provide explicit explanations about their decisions on errors [29]. A pilot test involving four participants was conducted to ensure correctness and clarity of experiment instructions and procedure. Minor research design modifications were made based on our observations and the results of the pilot study, as well as interviews with those participants.

Participants

Ten undergraduate students were recruited for this study from a mid-sized university on the east coast of the United States. They were all native English speakers and none of them was a professional editor. These participants were sophomores, juniors, and seniors from eight different majors. Sixty percent of the participants were female. They were told that the research study would take about two hours and each participant would be paid $20 for his or her participation.

Task

The experimental task mainly consisted of two parts—detecting speech recognition errors and providing possible explanations for every error detected. The goal of the error detection was to identify the discrepancy between words in the transcripts generated by a speech recognition system and the original speech rather than polish the language.

Eight transcripts were selected and presented in three different formats—text (actual speech recognition output)-only, text with alternative hypotheses, and text with both alternative hypotheses and background information. This helped us evaluate the merit of supplementary information to error detection. Alternative hypotheses included top-three alternative word hypotheses and top-nine alternative utterance hypotheses along with their confidence scores. Background information was represented with task scenarios that were originally used to elicit the speech. The transcript distribution for three different formats was 3:3:2. The text transcripts were randomly sorted for each participant to counterbalance the potential ordering effect. Moreover, transcripts for the background information setting were extracted from different task scenarios to avoid carryover effects.

Each participant was asked to provide more than one piece of evidence to support each of the identified errors. It was recommended for each participant to seek evidence from multiple perspectives without the assistance of specific CERD. This would allow us to acquire knowledge that was actually applied by participants to detect errors and, more importantly, that can help us discover new CERD.

Transcript Data

The transcripts were randomly extracted from a dictation corpus that was generated by a commercial speech recognition system under high-quality conditions from the spontaneous speech of 27 speakers [12, 39]. All of the speakers were native, but not professional, English speakers.

We used two criteria for selecting transcript data—recognition accuracy and transcript length. The recognition accuracy of the speech corpus, 84 percent, was chosen as the empirical error rate. A qualified text transcript should be neither too short to provide necessary context information for error detection, nor too long to be completed within the given amount of time. Based on our experience and observations, 90 words were selected as the target length. The descriptive statistics of selected transcripts are shown in Table 1. The two columns in the middle, "number of words in the output" and "number of words in the reference," represent actual speech recognition outputs and corresponding manually corrected reference transcripts, respectively. Each transcript was prepared in all three formats, as described in the previous section.

Procedure

Before their arrival, participants were asked to preview two online documents introducing error annotation schemes and examples of text transcripts and other related information (e.g., alternative hypotheses and confidence scores). During the experiment, each participant first took a pretest to assess his or her knowledge of the annotation schemes and comprehension of data that would be presented in the transcripts. If a participant made mistakes on one or more questions, he or she would be asked to redo the questions after reviewing the related online documents. Then, the participant moved on to read instructions and analyze the text transcripts in a given sequence. All the text transcripts were presented in hard copy workbooks, on which a participant could mark and annotate errors. In addition, the participant was asked to think and explain aloud by providing justification explanations whenever an error was detected, which were recorded with a digital voice recorder. After completing all the text transcripts presented in each formats, each participant was asked to fill out a short questionnaire about his or her perception of the process and the results of error detection.

A verbal protocol analysis relies on collecting information about the course and mechanisms of cognitive processes of the internal states of the problem solvers [11, 29]. The think-aloud instruction used in the experiment was: "Say out loud every justification that passed through your mind for explaining each error as you detect it."

Table 1. Statistics of the Selected Transcripts

Transcripts	Number of sentences	Number of words in the output	Number of words in the reference	Number of errors	Recognition accuracy
A	4	71	69	15	84.1
B	6	90	88	20	80.7
C	3	85	75	36	73.3
D	5	77	83	18	78.3
E	4	111	109	20	85.3
F	5	100	109	22	80.7
G	5	105	107	18	84.1
H	4	83	81	16	82.7

Participants were encouraged to think of more than one justification to acquire more thorough protocols. The contents obtained from the think-aloud method and immediate retrospective responses are valid.

Data Analyses and Results

IN THE FIRST STEP OF DATA ANALYSIS, verbal protocols recorded from each participant during the error detection process were segmented into the units of text transcripts. Content analysis was carried out on the segmented verbal protocols by two independent coders to interpret and encode them. An interrater comparison was conducted to validate the encoding results. Finally, CERD that emerged from the verbal protocol analysis were illustrated and their effectiveness was reported.

Data Encoding

Two coders were recruited to encode the recorded explanation independently. Their tasks were to listen to participants' explanations recorded in the audio files and interpret them while referring to the errors annotated in the completed workbooks. Based on their interpretation results, the coders filled out a coding worksheet in Excel (see Table 2) to record coding results separately. Particularly, coders were asked to fill out the parts (formatted in italics in Table 2) by specifying transcript# (A–H), sentence#, error#, error type (I: insertion; D: deletion; S: substitution), a list of cues (cue 1 and cue 2 were placeholders for specific cues), and their presence (i.e., *Y*) for specific errors. Such a process was repeated for every text transcript completed by each participant. The average independent coding time was five times of the audio length per text transcript per coder. Thus, the data encoding was a time-consuming and labor-intensive process.

Interrater Comparisons

To examine potential bias, the reliability of cues generated by the two independent coders was tested. There were a total of 1,275 errors for which cues were independently encoded by each coder. The employed think-aloud method was conducive to eliciting multiple cues for detecting each error from the participants. The number of cues per error ranged from 1 to 7, with a mean of 2.22 and a standard deviation of 1.0.

The text descriptions of cues in the encoding results were first normalized by reducing nouns to their singular forms, removing articles, transforming verb phrases to noun phrases, and so on. Then, all the cues were sorted in alphabetical order. The same cues were merged, and each distinctive cue was assigned with a unique identifier.

Given that there could be more than one cue category encoded for each error and some cues encoded by coder A did not appear in the list of cues encoded by coder B, and vice versa, Cohen's kappa, a popular measure for interrater reliability, was not appropriate for this type of qualitative data. Thus, we developed metrics to measure

Table 2. A Sample Coding Worksheet

Transcript#	*A*	*A*	. . .
Sentence#	*1*	*1*	. . .
Error#	*1*	*2*	. . .
Error type	*D*	*I*	. . .
Cue 1	*Y*		. . .
Cue 2		*Y*	. . .
. . .			. . .

the overall agreement on the separately encoded cues between coders A and B, which is shown in formulas (1) and (2) (or (1) and (2′)).

$$\text{overall agreement} = \frac{\sum_{p}\sum_{t}\sum_{s}\sum_{i}\text{consensus}\left(e_{p,t,s,i}\right)}{N} \tag{1}$$

$$\text{consensus}\left(e_{p,t,s,i}\right) = \begin{cases} 1 & \text{if } C_A^{p,t,s,i} \cap C_B^{p,t,s,i} \neq \Phi, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\text{consensus}\left(e_{p,t,s,i}\right) = \frac{\left|C_A^{p,t,s,i} \cap C_B^{p,t,s,i}\right|}{\left|C_A^{p,t,s,i} \cup C_B^{p,t,s,i}\right|}, \tag{2′}$$

where N is the total number of errors identified by all the participants, $e_{p,t,s,i}$ uniquely denotes error i identified by participant p in sentence s of transcript t, and $C_A{}^{p,t,s,i}$ (or $C_B{}^{p,t,s,i}$) denotes a set of cues encoded for error $e_{p,t,s,i}$ by coder A (or B). It is noted that two alternative formulas were provided for computing the consensus on the coding results for $e_{p,t,s,i}$. Formula (2) concerns whether there are any overlapping cues to $e_{p,t,s,i}$ between two coders, whereas formula (2′) takes into account the percentage of cues shared by the two coders. The latter is more conservative than the former.

The resulting overall agreement was about 62.0 percent and 43.5 percent based on formulas (2) and (2′), respectively. The relatively low agreement rates were attributable to several reasons: (1) each coder had his or her own preference in choosing words to represent specific cues, (2) the similarity between cues that were represented at different levels of granularity were not factored into the comparison, and (3) different cues were encouraged because the goal of this research was to discover new knowledge from users rather than validate existing knowledge. After mapping leaf node cues to an upper level according to the hierarchy of CERD, which will be introduced in the next section, the overall intercoder agreement reached 71.5 percent and 55.0 percent based on formulas (2) and (2′), respectively. Given the complexity of the encoding situation, especially the large number of categories and the high frequency of multicategory encodings, we believe that the agreement rates were reasonable.

The disagreements in the encoded cues for individual errors are further investigated by the first author and addressed by either consolidation or discussion with the coders. The following two scenarios were resolved by consolidation:

- If two cues had overlap, the nature of their relationship was further analyzed. For example, if it was a generic–specific relationship (e.g., incompatible semantics between two constituents versus incompatible semantics between subject and predicate), the more generic cue would be discarded.
- If two cues were disjointed and contradictory with each other (e.g., part-of-speech confusion versus open-class word choice), the corresponding verbal protocol would be examined by the first author and one of the cues would be eliminated.

For the remaining inconsistent results, two coders would revisit the original audio files, discuss their discrepancy under the facilitation of the first author, and reach a consensus or choose the majority opinion. As a result, cues that were found to be complementary to each other were kept, different expressions of the same cues were merged, and cues resulting from the coders' misinterpretation were dropped. Finally, 53 cues were selected for the final list of CERD.

A Taxonomy of CERD

An analysis of the discovered CERD revealed that some of cues were relevant to each other. Drawing upon the literature from linguistics, natural language processing, and speech recognition, we clustered different CERD based on the closeness of their relationships and developed a taxonomy of CERD using the bottom-up approach. The top two levels of the CERD hierarchy are shown in Figure 1. CERD were first grouped into three categories—linguistics-based, hypotheses-based, and others. The linguistics-based CERD were further divided into phonological, morphological, syntactic, semantic, and discourse types. The hypotheses-based CERD included both word and utterance hypotheses. Others type of CERD contained adjacent errors, language style, and unnecessary repetitions.

Each leaf node in Figure 1 can be further expanded into subcategories. The full taxonomy of CERD is provided in the Appendix. For the sake of space, we only used the syntactic node for illustration purpose. According to the fully expanded hierarchy shown in Figure 2, the syntactic CERD consist of phrase structure and sentence structure, and the phrase-structure CERD are further decomposed into parallel structure, modifier, and so on.

To uniquely identify CERD in the taxonomy, we represented the top two levels of CERD with boldface letters in Figure 1 and delineated the lower levels of CERD with sequential numbers in Figure 2. Moreover, the identifier of a CERD at a lower level is created by attaching the corresponding delineating letter (or number) to the identifier of its parent CERD using a period as the delimiter. For example, "LG" stands for linguistics-based syntactic CERD and LG.1.2 stands for "parallel structure," where "1" denotes the first child of LG, and "2" denotes the second child of the phase-structure

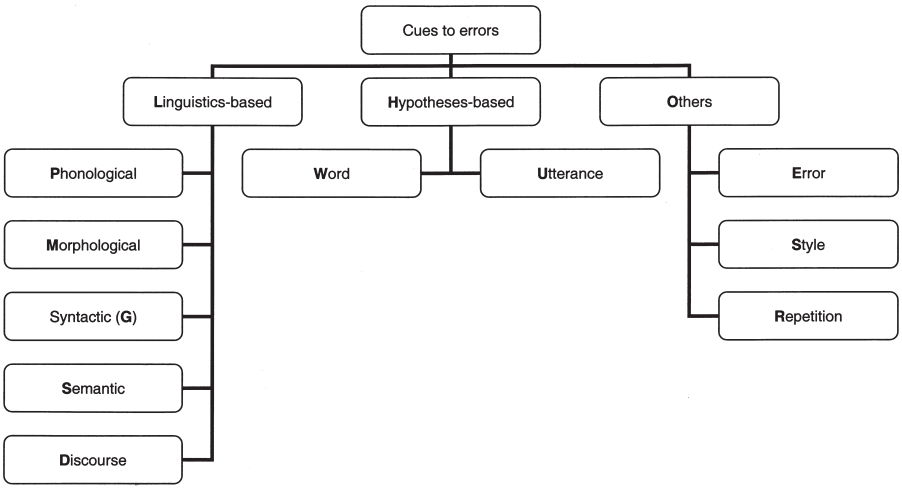

Figure 1. A Taxonomy of CERD (Top Two Levels)

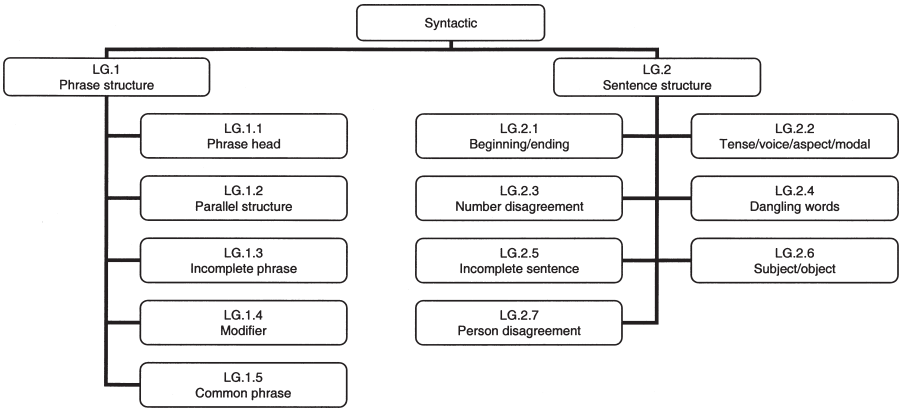

Figure 2. The Hierarchy of Syntactic CERD

CERD. Due to space constraints, the rest of the discussion focuses on the CERD in the third layer. Detailed descriptions and examples of these CERD are given in Table 3. Since all the language-style CERD were mainly induced by user-introduced rather than system-generated errors, they were grouped together and referred to as OS in Table 3.

Descriptive Statistics of CERD

Our analysis of the encoding results showed that some CERD were used more frequently than others. To evaluate the effectiveness of individual CERD, we needed metrics that could take two factors into consideration: (1) the measurement should be positively proportional to the usage frequency of CERD, and (2) the measurement

Table 3. Descriptions and Examples of CERD

IDs	CERD	Descriptions	Examples[1]
HU.1	Utterance hypotheses	Alternative hypotheses at the utterance level. (An utterance is a continuous acoustic sequence.)	{{easier than I had expected} {14 –6 –5 5 40}} {{easier than I have expected} {14 –6 –5 3 40}}
HU.2	Utterance length	The length of utterance in which an output word appears.	{{easier than I had expected} {14 –6 –11 5 40}}
HU.3	Path ratio	Ratio between number of paths containing a word and total number of paths.	The lengths of the above three utterance hypotheses are equal to 5, and the path ratio of "had" is 67 percent (2/3).
HW.1	Word hypotheses	Alternative word hypotheses.	{{quickly} 33} {{quietly} 9} {{quirkily} 1}
HW.2	Confidence scores	Word confidence scores.	There are three alternative word hypotheses in
HW.3	Highest confidence score	Whether a word has the highest confidence score among all the alternatives.	the above list. The confidence score of "quickly" is 33, which is the highest among the three.
LD.1	Out of context	Incompatibility between the current word and the surrounding discourse.	This game is likely held in *inland* because that is where John is from. . . . Also, John Wild lived in London for most of his life.
LD.2	Coreference	Inconsistency between the reference with the referent in the discourse.	Even though I do not know very much about *truck,* I feel that I have a good idea of what kind of employee he is.
LD.3	Contradictory information	Contradiction between the meaning of a word and the discourse.	I have many questions and often *can* decide between two conflicting ideas and he was one of the few people always available to give ideas.
LD.4	Background scenario	Background scenarios used to elicit the original speech.	Describe a book that you are going to write.

(continues)

Table 3. Continued

IDs	CERD	Descriptions	Examples[1]
LG.1	Phrase structure	Ungrammatical phrase structure, including phrase head, parallel structure, incomplete phrase, modifier, and common term.	Even though Paul received his Ph.D. in computer science, he is really interested in *sell* issues of information systems.
LG.2	Sentence structure	Ungrammatical sentence structure, including illegal beginning/ending of a sentence, number/tense/voice/ aspect/modal/person disagreement, dangling words, subject, object, or incomplete sentence.	He *pick* up the kids from school everyday. I reserved a [2] for my mom for a Broadway show.
LM.1	Part-of-speech confusion	Misuse of a word of a different part-of-speech (e.g., determiner, preposition, pronoun, conjunction, and verb).	The fact that they all have common subjects makes them *close* related.
LM.2	Open-class word choice	The choice of a wrong content word such as noun, verb, adjective, or adverb.	I am able to *add* very personable with clients and develop a trusting relationship with them.
LM.3	Closed-class word choice	The choice of a wrong function word from parts-of-speech such as proposition and determiner.	I would like to continue working on this product *and* the coming year.
LM.4	Nonword choice	Problem with choosing punctuation marks, including missing, extraneous, or wrong punctuation marks such as commas and parentheses.	I would like to give feedback about Susan and *(* (this name is fictitious because . . . would indicate the specific individual).
LP.1	Disfluency	Speech disfluency such as false start, repetitions, repairs, and filler words.	I went to medical school in 1970s but *a* I did not finish it.
LP.2	Phonetic similarity	A word mistakenly used due to sharing similar sound to another word.	Peter is a very sociable *parson,* however, he tries to focus on too many things at the same time.
LP.3	Word split	One word being split into more than one word.	The conference that I would like to attend is the World Financial conference held in London *and when.*

LS.1	Making no sense	A word does not make sense in the sentence.	In particular, Janet *kilns* and *move* will be traveling with me.
LS.2	Incompatible semantics	Incompatibility in the meaning between two sentence constituents, including subject–object, preposition–object, modifier–head, subject–predicate, and predicate–object.	I could purchase a new *pressure.* Some of *diary* food must be refrigerated such as eggs that can go *back.*
OE.1	Preceding error	The preceding word is an error.	In reviewing one of my peers, *and cited* to choose Peter because it was only until this week that I really got to know him.
OE.2	Following error	The following word is an error.	The travel arrangements [2] *couldn't* for this conference is I would like to take a cruise liner from Baltimore to London.
OR.1	Redundancy	Repeat the same word more than once in a row.	I was very happy to *to* meet an old friend who just visited Baltimore.
OS	Language style	Lack of conformance to traditional writing style such as using wrong letter cases, contractions, and double negations.	*it* causes my entire paper to be lost.

Notes: [1] words shown in italics are recognition errors; [2] a deletion error.

should reflect the chance that certain CERD were applicable. For example, alternative hypotheses were available in the alternative hypotheses format but not in the text-only format in our experiment. As a result, we adapted the concept of support from the association rule mining to measure the effectiveness of CERD. Support was defined as the ratio of the number of errors for which a CERD was actually used divided by the total number of errors for which that CERD could be used. The levels of support for the CERD are reported in Table 4 in descending order.

There were two types of support values—all and correct. The former was based on all the detected errors, whereas the latter was based on the errors detected correctly. It is shown in Table 4 that the rankings of CERD appeared to be consistent between both types of support for the majority of CERD except OS.3 and LM.4, which were ranked much lower for *correct* support than for *all* support. This is because we did not consider letter case and punctuation as the sole explanation for a recognition error. Therefore, discussion in the next section focuses on all support.

Discussion

OUR IMPLEMENTATION OF THE USER-CENTERED APPROACH to discovering CERD resulted in a taxonomy of CERD. The taxonomy provides broad implications and benefits to both research and practice in error detection, including generic error detection, context-sensitive error detection and correction, and knowledge support for error detection.

Implications to Generic Error Detection

The taxonomy allows researchers to investigate CERD in a systematic way and guides future research and practice in error detection. As shown in Table 4, LS.1 (making no sense) received the highest support among all the CERD. LS.2 (incompatible semantics) was also well supported. The former represents the cases in which a word does not fit in a sentence or is irrelevant to the meaning of the sentence. The latter represents semantic mismatches between two sentence constituents (e.g., subject–object and modifier–head). Both CERD suggest that semantic analysis, a process to determine what each word means and what a sentence means when individual words are combined with each other, is crucial to detecting semantic anomalies caused by speech recognition errors. Moreover, word co-occurrence analysis could be helpful to determine whether or not a word makes sense in a sentence.

LG.1 (phrase structure) and LG.2 (sentence structure) were next after LS.1 in Table 4. They suggest that syntactic information, including parallel structures; modifier–header relationships; sentence completeness; and number-, person-, tense-, and voice-agreements; and so on, is indispensable to detecting speech recognition errors. Whether or not a word can be parsed and the probability of a word to be parsed are good syntactic/semantic indicators [37, 49]. The state-of-the-art technologies for syntactic parsing

Table 4. CERD and Their Levels of Support

IDs	CERD	Support		IDs	CERD	Support	
		All	Correct			All	Correct
LS.1	Making no sense	0.316	0.360	LP.2	Phonetic similarity	0.021	0.026
LG.1	Phrase structure	0.259	0.263	LD.4	Background scenario	0.019	0.022
LG.2	Sentence structure	0.164	0.186	HU.1	Utterance hypotheses	0.018	0.019
HW.1	Word hypotheses	0.154	0.180	HU.2	Utterance length	0.015	0.015
HW.2	Confidence scores	0.149	0.175	OE.1	Preceding error	0.013	0.008
LM.1	Part-of-speech confusion	0.101	0.121	LP.3	Word split	0.012	0.015
LS.2	Incompatible semantics	0.082	0.102	LM.4	Nonword choice	0.012	**0.003**
HW.3	Highest confidence score	0.080	0.092	OS.4	Misplacement	0.009	0.006
LP.1	Disfluency	0.061	0.059	OE.2	Following error	0.006	0.004
OS.3	Letter case	0.057	**0.013**	LD.2	Coreference	0.005	0.006
LD.1	Out of context	0.054	0.057	HU.3	Path ratio	0.004	0.005
LM.2	Open-class word choice	0.051	0.058	OS.2	Double negation	0.002	0.003
OR.1	Redundancy	0.046	0.032	LD.3	Contradictory information	0.002	0.002
LM.3	Closed-class word choice	0.035	0.029	OS.1	Contraction	0.001	0.000

Note: Boldface figures refer to lowered rankings for the CERD when switching from all support (based on all the detected error) to correct support (based on correctly detected errors only).

are relatively mature, which could play a significant role in the detection of recognition errors.

HW.1 (hypotheses), HW.2 (confidence scores), and HW.3 (highest confidence score) were among the best supported hypotheses-based CERD. They reveal that, in general, alternative word hypotheses and their confidence scores are helpful to users in detecting speech recognition errors. They can be used to discriminate top hypotheses and to infer that a word is possibly wrong if its confidence score is much lower than those of other alternative words [44]. This implies that enhancing confidence measures is important to improving error detection.

LM.1 (part-of-speech confusion) follows HW.2 in terms of the level of support. Part-of-speech confusion may lead to ill-formed sentences. Consequently, traditional parsing technologies become less effective in dealing with sentences containing such errors and can even fail to produce an output. Therefore, a robust natural language parser that can indicate where a sentence breaks would be extremely valuable.

Among all the phonological CERD, LP.1 (disfluency) received the highest support. During spontaneous dictation, a participant must plan for the next sentence while speaking. Thus, disfluencies such as false start become useful CERD to identify errors in continuous speech such as dictation, meeting conversation, and presentation. The approaches to automatic disfluency detection have utilized prosodic information, statistical word language models, syntactic structures, textual information, and lexical features [24]. Nevertheless, this line of research is still at an early stage. Given that acoustic processing is insufficient to solve disfluencies, postprocessing would be necessary to reduce the ambiguity inherent in a single knowledge source. For example, repairs and false start may be detected via linguistic means.

OS.3 (letter case) was best supported among others type of CERD. Many speech recognizers either generate words only in lowercase or only partially address the letter case problem. This is largely attributed to the difficulty associated with detecting sentence boundaries. Although, in this study, the support of letter case was much lower for correct detection than all detection, it is shown that the letter case affects the usability of a speech recognition system. For example, it is uncommon to start a sentence with a lowercase and capitalize some common words within a sentence. Such problems may be addressed by sentence boundary identification and name entity recognition techniques, respectively.

LD.1 (out of context) was the best supported discourse-level CERD. It indicates that other adjacent sentences provide a useful context for detecting errors in the current sentence. Sometimes it is necessary to look for useful CERD beyond one sentence. Unlike domain-specific dialogue systems or task-oriented recognition systems, the discourse information obtained from large vocabulary dictation recognition is less structured and predictable. Nonetheless, models and theories on the rhetoric structure of text [28] from discourse processing and computational linguistics fields can guide the effort on incorporating contextual information into error detection.

It is shown in Table 4 that OR.1 (redundancy) turned out to be a concern for a speech recognition system. This is partly due to the nature of dictation, in which a speaker tends to repeat previous words when a recognition system does not transcribe

them immediately or when the output is wrong. Moreover, repetition of the same punctuation marks in a consecutive sequence in dictation appeared to be atypical to the participants.

LM.2–LM.4 refer to choosing a wrong (1) function word (e.g., conjunction), (2) content word (e.g., noun), or (3) nonword string (e.g., punctuation), respectively. The first two were supported better than average, but the last one was not supported well. In all three cases, a wrong word was chosen from the same grammatical category or a wrong punctuation mark was selected. Different techniques are required to deal with different kinds of words. For example, to detect the problem of wrong content word choice, semantic or discourse processing could help examine whether the word fits in a phrase, a sentence, or even a discourse. In addition, the problem of wrong function word choice may be addressed with corpus-based analysis of the collocation of function words and content words or dictionary-based pattern matching.

Those CERD whose levels of support were greater than 0.01 and less than 0.03 included LP.2 (phonetic similarity), LD.4 (background scenario), HU.1 (utterance hypotheses), HU.2 (utterance length), OE.1 (preceding error), and LP.3 (word split). Differentiating homonyms consisting of similar phones is a fundamental challenge in speech recognition. The problem of phonetic similarity is exacerbated in recognition errors that cut across word boundaries. This is rooted in incorrect segmentation of acoustic signal sequence of continuous utterance as well as *n*-gram language models that are widely adopted in speech recognition engines. A speech recognition error may generate a domino effect and result in further errors in subsequent words [7]. It implies that correctly detecting one recognition error may lead to the successful detection of other errors in adjacent words. In our study, background scenarios were found to play a notable role in error detection by providing discourse information about the text transcript, which can be especially useful in situations where the third party is involved. Like word hypotheses, utterance hypotheses and utterance length can also signal recognition errors. Longer utterances provide more context for detecting an erroneous word.

The next group of CERD whose support was at least 0.005 but no more than 0.01 included OS.4 (misplacement), OE.2 (following error), and LD.2 (coreference). The remaining CERD that were least supported in this study included HU.3 (path ratio), OS.2 (double negation), LD.3 (contradictory information), and OS.1 (contraction). Most of the above CERD deal with writing style issues or involve deep understanding of discourse or hypotheses information, and thus were not well supported.

Implications to Context-Sensitive Error Detection and Correction

Intuitively, different types of speech recognition errors (i.e., deletion, insertion, and substitution) require different CERD. Possible associations between error types and CERD can support context-aware error detection and correction by users or systems. On one hand, if specific CERD are applicable to certain words, they will not only facilitate detecting possible errors but also suggest how to correct errors. On the other

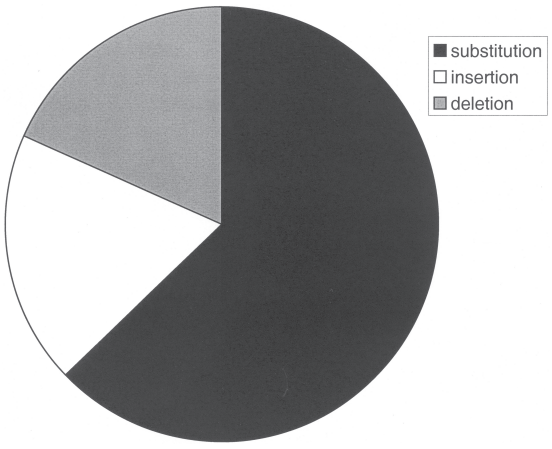

Figure 3. Distribution of Error Types

hand, once an error type is specified, error correction can be better guided by a recommended subset of relevant CERD.

An analysis of error distribution in the text transcripts used in this study showed that substitution, insertion, and deletion errors accounted for 62.9 percent, 18.3 percent, and 18.8 percent, respectively, as shown in Figure 3. This distribution was very close to the error distribution in the entire speech corpus (substitution: 63.7 percent; deletion: 18.7 percent; insertion: 17.6 percent). Apparently, substitution was the predominant type of error.

The error type distribution is echoed in the distribution of CERD in terms of error types, as shown in Table 5. For example, CERD such as HU.3, LD.2–LD.4, LM.2, LP.3, and OS.1–OS.3 were used solely to detect substitution errors; CERD such as HU.1, HW.1, HW.3, LM.3, LP.2, and LS.2 were used predominantly (> 80 percent) to detect substitution errors. These findings suggest that hypotheses-based, style-related, and most discourse CERD are particularly useful for detecting substitution errors. They also confirm our anticipation that CERD concerning word choice issues (e.g., LM.2 and LM.3), phonetic similarity (i.e., LP.2), and semantic incompatibility (i.e., LS.2) are effective for handling substitution errors.

Some CERD, including HU.2, HW.2, LD.1, LS.1, and OS.4, were used primarily for substitution errors and occasionally for insertion errors. This corroborates the previous finding that hypotheses-related CERD are helpful to detect substitution errors. In addition, when an output word makes no sense or is perceived to be out of context or misplaced, the word can be corrected mostly likely by replacing it with a different word or sometimes by deleting it. LG.1, LG.2, and LM.1 were mostly used to address substitution errors and sometimes deletion and insertion errors.

It was interesting to observe that, although both OE.1 and OE.2 were associated with substitution and deletion errors, OE.1 was more commonly used in detecting substitution errors and OE.2 was more popular in detecting deletion errors. The former

Table 5. Results of Association of CERD with Error Types

Associations*	Substitution	Insertion	Deletion
Strong (ε > 60 percent)	HU.3, LD.2, LD.3, LM.2, LP.3, LD.4, OS.1, OS.2, OS.3 HU.1, HW.1, HW.3, LM.3, LP.2, LS.2 HU.2, HW.2, LD.1, LS.1, OS.4 LG.1, LG.2, LM.1	LM.4, OR.1	OE.2
Medium (ε > 30 percent)	OE.1, LP.1	HU.2, HW.2, LD.1, OS.4, LS.1, LP.1	LM.4, OE.1
Weak (ε > 10 percent)	OE.2	LG.1, LG.2, LM.1, OE.1	LG.1, LG.2, LM.1, LP.1

* ε is the probability of the CERD used for the error type.

was also occasionally used for insertion errors. Insertion errors accounted for the majority of the cases handled with LM.4 (nonword choice) and OR.1 (redundancy). As a result, nonwords and redundant words are most likely to be removed in error correction. LM.4 was also used for deletion errors occasionally to indicate missing punctuation. LP.1 was more likely to pinpoint substitution and insertion errors than deletion errors. This indicates that the output due to disfluency noise is likely to be corrected by word replacement or elimination.

In sum, more CERD were strongly associated with substitution errors than insertion and deletion errors. Since many CERD tend to be associated with a specific error type, the analysis of the relationship between different CERD may provide additional insights into error detection, which will be discussed in the next section.

Implications to Knowledge-Supported Error Detection

By analyzing the associations between different CERD, we will be able to empower users or systems with knowledge to help them make more informed and better judgments in error detection. This will also lead to the development of knowledge-based support systems [31] user error detection.

We used the Jaccard coefficient [1], a popular metric for query-document matching in information retrieval, to measure the association between different CERD mainly for two reasons: (1) the Jaccard coefficient measures asymmetric information on binary variables (CERD are represented as binary variables), and (2) a pair of CERD usually co-occur in a small number of errors and co-absent in a large number of errors. Unlike traditional correlation coefficients, the Jaccard coefficient helps remove double-absence cases that make little contribution to the association between two CERD. The results are reported in Table 6. Based on our observation, we empirically selected 0.05 as the threshold to distinguish strong and weak associations. The OS type of CERD was excluded from the analysis because those CERD addressed user errors rather than recognition errors.

As shown in Table 6, LS.1 was most active, with strong associations with many other CERD such as LD.1, LG.1, LM.1, LM.2, LS.2, HW.1, HW.2, and HW.3. When an output word is out of context, has incompatible semantics, or results in ill-formed phrase structures, it is unlikely the word will make any sense in a sentence. Choosing a wrong word from the same or a different grammatical category can suggest that the word does not make sense in the sentence. Moreover, alternative word hypotheses-related CERD can facilitate the judgment of whether a word fits in a sentence or not. In this study, LS.2 also played an active role in error detection, and it was highly correlated with LM.1 and LM.2. In addition, word hypotheses and whether they have the highest confidence scores can provide information about the semantic compatibility of an output word. Therefore, the judgment of whether a word makes sense or not can be enhanced by using other types of linguistic information such as discourse, syntactic, and morphological CERD and hypotheses.

LG.1 was highly correlated with LM.1, LP.1, HW.1, and LS.1. Specifically, ill-formed phrase structures may be caused by wrong parts-of-speech, inserting/deleting

Table 6. Results of Associations Between CERD

	HU.1	HU.2	HU.3	HW.1	HW.2	HW.3	LD.1	LD.2	LD.3	LD.4	LG.1	LG.2
HU.2	0.040											
HU.3	**0.063**	**0.071**										
HW.1	0.015	0.007	0.000									
HW.2	0.015	0.016	0.008	**0.169**								
HW.3	0.000	0.013	0.031	**0.154**	**0.096**							
LD.1	0.012	0.000	0.000	0.043	0.039	0.000						
LD.2	0.000	0.000	0.000	0.008	0.000	0.000	0.027					
LD.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				
LD.4	0.000	0.000	0.000	0.000	0.008	0.000	0.027	0.000	0.000			
LG.1	0.000	0.006	0.000	**0.068**	0.044	0.029	0.028	0.000	0.000	0.012		
LG.2	0.005	0.000	0.005	**0.064**	0.048	0.042	0.018	0.000	0.000	0.009	0.049	
LM.1	0.000	0.007	0.000	**0.073**	**0.051**	**0.055**	0.021	0.008	0.000	0.000	**0.093**	**0.073**
LM.2	0.026	0.000	0.000	0.044	0.028	0.024	**0.055**	0.000	0.000	0.043	0.023	0.019
LM.3	0.000	0.000	0.000	0.018	0.012	0.000	0.009	0.000	0.000	0.000	0.014	0.012
LM.4	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000
LP.1	0.000	**0.071**	0.000	**0.081**	0.026	0.036	0.007	0.000	0.000	0.000	**0.054**	0.032
LP.2	0.000	0.026	0.000	0.020	0.035	0.000	0.000	0.000	0.000	0.000	0.017	0.004
LP.3	**0.074**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LS.1	0.005	0.005	0.002	**0.069**	**0.074**	**0.052**	**0.098**	0.015	0.005	0.005	**0.083**	0.041
LS.2	0.000	0.009	0.000	**0.066**	0.028	**0.070**	0.018	0.009	0.000	0.000	0.036	0.013
OE.1	0.033	0.036	0.000	0.007	0.007	0.025	0.024	0.000	0.000	0.043	0.015	0.004
OE.2	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.000	0.000	**0.077**	0.006	0.005
OR.1	0.000	0.000	0.000	0.006	0.017	0.000	0.000	0.000	0.000	0.000	0.003	0.004

(continues)

Table 6. Continued

	LM.1	LM.2	LM.3	LM.4	LP.1	LP.2	LP.3	LS.1	LS.2	OE.1	OE.2
HU.2											
HU.3											
HW.1											
HW.2											
HW.3											
LD.1											
LD.2											
LD.3											
LD.4											
LG.1											
LG.2											
LM.1											
LM.2	0.000										
LM.3	0.000	0.000									
LM.4	0.000	0.000	0.000								
LP.1	0.020	0.021	0.000	0.000							
LP.2	0.047	**0.057**	0.000	0.000	0.000						
LP.3	0.000	**0.096**	0.000	0.000	0.000	0.000					
LS.1	**0.054**	**0.066**	0.018	0.002	0.026	0.021	0.000				
LS.2	**0.050**	**0.070**	0.042	0.007	0.022	0.040	0.008	**0.059**			
OE.1	0.007	0.012	0.000	0.000	0.033	0.000	0.000	0.007	0.000		
OE.2	0.000	0.014	0.000	0.000	0.024	0.000	0.000	0.005	0.000	**0.200**	
OR.1	0.000	0.008	0.000	0.000	0.030	0.000	0.000	0.013	0.000	0.000	0.000

Note: Boldface figures refer to strong associations (> 0.05) between two CERD.

a word, or speech disfluency. Alternative word hypotheses can also be a point of reference for possible problems with phrase structures. For similar reasons, problems with sentence structures (LG.2) can be substantiated with morphological (e.g., LM.1) and word hypotheses (i.e., HW.1) information.

LM.1 was highly associated with HW.1–HW.3 and some semantic and syntactic CERD (e.g., LS.1, LS.2, LG.1, and LG.2). Alternative word hypotheses may contain the correct word, indicating possible confusion in the part-of-speech of a recognized word.

In addition to semantic CERD (e.g., LS.1 and LS.2), LM.2 was also highly associated with some phonological (e.g., LP.2 and LP.3) and discourse CERD (e.g., LD.1). The results suggest that a misrecognized content word is likely to sound similar to, or be a part of, the content word being spoken or to be out of context.

The strong associations between different word hypotheses CERD (e.g., HW.1–HW.3) in Table 6 show that they mutually reinforce each other. As discussed above, word hypotheses and related information are strongly associated with some morphological, syntactic, and semantic cues (i.e., LM.1, LG.1, LG.2, LS.1, and LS.2). Furthermore, HW.1 was also found to correlate strongly with LP.1, implying that word hypotheses are suggestive of disfluency.

Like word hypotheses, utterance hypotheses CERD were highly correlated with one another except for the relatively weak relationship between HU.1 and HU.2. The results imply that HU.3 (path ratio) complements other utterance hypotheses CERD. The high correlation between HU.1 and LP.3 suggests that if two consecutive words are perceived to be the outcome of an inappropriate split of one word, there is a good chance to catch such an error by referring to alternative utterance hypotheses. It is interesting to find that HU.2 has a strong association with LP.1, which warrants a further examination of the relationship between utterance length and disfluency.

Finally, our analysis revealed that OE.1 and OE.2 were highly correlated with each other. It infers that if there is an error with the preceding word, it would be helpful to check the following word in judging the correctness of the current word, and vice versa. This highlights the phenomenon of consecutive errors in the recognition output. Moreover, if the previous word is an error, background information (LD.4) can be scrutinized to help determine if the current word is possibly wrong.

By applying the knowledge acquired from multiple correlated sources, we should reduce the ambiguity associated with a single knowledge source in error detection.

Summary of CERD and Discussion

Using a user-centered approach, this study discovered a variety of CERD, including morphological, syntactic, semantic, hypotheses-related, and others, which were useful in detecting recognition errors. Some CERD received the best support in this study, including LS.1 (making no sense), LG.1 (phrase structure), LG.2 (sentence structure), HW.1 (word hypotheses), HW.2 (confidence scores), and LM.1 (part-of-speech confusion). Moreover, the utility of CERD was examined in relation to error types and other CERD. By incorporating CERD into an error correction system, automatic

error detection and correction can be improved. The findings of this research also collectively lay the foundation for developing a decision support system that is able to recommend a group of correlated and context-sensitive CERD to facilitate users in detecting speech recognition errors.

The developed taxonomy of CERD not only provides new evidence to support some findings of previous studies but also discovers some promising new CERD to guide future error detection. For example, word hypotheses have been explored in error detection or correction [20, 43, 44, 45, 49]. The highest confidence score, approximated by the difference in the confidence scores between the best and the second-best hypotheses, was used to develop confidence measures [44]. In addition, the second-best hypothesis was used by some systems to replace the best hypothesis that was possibly wrong in error correction [27]. Preceding error echoes the observations of other studies [7, 12, 17].

Some linguistic CERD, such as LM.1, LG.1, and LG.2, were incorporated as features to derive the scores of confidence measures and showed promising results [37, 41, 49]. Open-class word choice was used to detect errors at the word level [41]. However, the encoding of syntactic information in previous studies was restricted to either a probabilistic structured language model [37] or a probabilistic or nonprobabilistic representation of whether a word can be parsed [37, 49]. It is rare to encode specific syntactic knowledge explicitly. Such knowledge is found to be beneficial to error detection in this study.

Semantics-induced CERD were suggested by several other studies. For example, co-occurrence analysis was used to detect words that were incompatible with the surrounding words [38]. Features extracted by a semantic parser were helpful to exposing semantic incompatibility among different components [49]. The discourse CERD were incorporated into a dialogue system to indicate whether a word was already mentioned in the previous dialogue [41]. Co-occurrence analysis could also make use of long-range contextual features beyond those in the current sentence [38]. Nonetheless, this study represents the first effort to specify concrete types of semantic incompatibility and discourse incongruity caused by speech recognition errors.

Some CERD that were selected by previous studies did not emerge from this study. For example, path ratio [7, 15, 36, 44, 49, 51] is a complex measure, which may not be intuitive for general users to employ immediately.

It is encouraging to discover some new CERD in this study. For example, word split, coreference, background scenario, and redundancy are promising CERD that have potential in speech error detection and correction. Moreover, some syntactic and semantic CERD obtained from the experiment remain to be fully explored in improving the speech recognition output. The syntactic and semantic information that has been previously applied is restricted to whether words "can be parsed" or "co-occurred." Semantic and pragmatic knowledge can potentially address a key factor liable for recognition errors—lack of commonsense understanding of what is being said [8]. Therefore, automatic error detection will greatly benefit from the advance of natural language processing. Meantime, challenges faced in improving related linguistic techniques, such as coreference resolution, should never be underestimated.

Verbal protocol analysis, as we conducted in this study, was both time-consuming and labor intensive. Participants' data had to be reviewed several times by each encoder in order to avoid possible misinterpretation of their intentions. Moreover, reexamination was followed to consolidate and normalize the encoding results for different participants and to resolve disagreements between coders. We hope that the taxonomy created in this study provides a jump-start and general guidance for future studies along this line.

The findings on CERD in this study were based on dictation speech recognition. They do not depend on dictation but can be extended to other applications of speech recognition. Nevertheless, the effectiveness of CERD may vary as the speech context changes. For example, sentence hypotheses may become an important feature in dialogue recognition. As a result, the CERD reported here need to be reevaluated in other types of speech applications.

Conclusion

DETECTING AND CORRECTING RECOGNITION ERRORS are important and challenging issues in achieving widespread adoption of speech technology. In this research, we embarked on a quest to discover CERD in speech recognition output using a user-centered approach. A taxonomy of CERD was created based on content analysis of verbal protocols collected from a user experiment. The findings of this study can guide future research efforts to improve recognition output and aid users in detecting speech recognition errors.

This research makes multifold contributions to improving the usefulness of speech technologies. First, the developed CERD taxonomy is the first taxonomy in this field, which not only advances our knowledge on CERD but also provides a systematic organization of CERD for future reference. Second, this is the first study to apply a user-centered approach to discovering CERD, which overcomes the limitations of and complements the traditional data-driven approach. Third, we propose a new measure (i.e., support) to assess the effectiveness of CERD. Fourth, to the best of our knowledge, this is the first attempt to analyze CERD in relation to error types (i.e., substitution, insertion, and deletion). These findings, coupled with the ability to learn from users' behavior during error correction, will enable the development of knowledge-based, context-aware, and personalized systems to ease users' effort in error detection and correction.

The current work can be extended in several directions, such as implementing CERD and assessing their validity for automatic error detection, evaluating the impact of CERD on the user's performance in error detection, and investigating and comparing CERD in other types of speech applications such as dialogue systems.

Acknowledgments: This material is based upon work supported by the National Science Foundation under grant number 0328391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Anderberg, M.R. *Cluster Analysis for Applications.* New York: Academic Press, 1973.

2. Arnold, S.C.; Mark, L.; and Goldthwaite, J. Programming by voice, VocalProgramming. In M. Tremaine, E. Cole, and E. Mynatt (eds.), *Proceedings of the Fourth International ACM Conference on Assistive Technologies.* New York: ACM Press, 2000, pp. 149–155.

3. Bain, K.; Basson, S.H.; and Wald, M. Speech recognition in university classrooms: Liberated learning project. In V.L. Hanson and J.A. Jacko (eds.), *Proceedings of the Fifth International ACM Conference on Assistive Technologies.* New York: ACM Press, 2002, pp. 192–196.

4. Brill, E.; Florian, R.; Henderson, J.C.; and Mangu, L. Beyond *n*-grams: Can linguistic sophistication improve language modeling? In C. Boitet and P. Whitelock (eds.), *Proceedings of the Thirty-Sixth Annual Meeting on Association for Computational Linguistics.* Morristown, NJ: Association for Computational Linguistics, 1998, pp. 186–190.

5. Carpenter, P.; Jin, C.; Wilson, D.; Zhang, R.; Bohus, D.; and Rudnicky, A.I. Is this conversation on track? In P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan (eds.), *Proceedings of the Seventh European Conference on Speech Communication and Technology.* Bonn, Germany: International Speech Communication Association, 2001, pp. 2121–2124.

6. Chase, L. *Error-Responsive Feedback Mechanisms for Speech Recognizers.* Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1997.

7. Chase, L. Word and acoustic confidence annotation for large vocabulary speech recognition. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (eds.), *Proceedings of the Fifth European Conference on Speech Communication and Technology.* Bonn, Germany: International Speech Communication Association, 1997, pp. 815–818.

8. Deng, L., and Huang, X. Challenges in adopting speech recognition. *Communications of the ACM*, *47,* 1 (January 2004), 69–75.

9. Duchateau, J.; Demuynck, K.; and Wambacq, P. Confidence scoring based on backward language models. In F.J. Taylor, J. Principe, and H. Bourlard (eds.), *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2002, pp. 221–224.

10. Ein-Dor, P., and Spiegler, I. Natural language access to multiple databases: A model and a prototype. *Journal of Management Information Systems*, *12*, 1 (Summer 1995), 171–197.

11. Ericsson, K.A., and Simon, H.A. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press, 1993.

12. Feng, J., and Sears, A. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Transactions on Computer–Human Interaction*, *11*, 4 (December 2004), 329–356.

13. Furui, S. Automatic speech recognition and its application to information extraction. In R. Dale and K. Church (eds.), *Proceedings of the Thirty-Seventh Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ: Association for Computational Linguistics, 1999, pp. 11–20.

14. Gauvain, J.-L., and Lamel, L. Large vocabulary speech recognition based on statistical methods. In W. Chou and B.H. Juang (eds.), *Pattern Recognition in Speech and Language Processing*. Boca Raton, FL: CRC Press, 2003, pp. 149–189.

15. Gillick, L.; Ito, Y.; and Young, J. A probabilistic approach to confidence estimation and evaluation. In M.K. Lang and H. Hoge (eds.), *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2. Los Alamitos, CA: IEEE Computer Society Press, 1997, pp. 879–882.

16. Hagen, A.; Connors, D.A.; and Pellom, B.L. The analysis and design of architecture systems for speech recognition on modern handheld-computing devices. In R. Gupta and Y. Nakamura (eds.), *Proceedings of the First IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis.* New York: ACM Press, 2003, pp. 65–70.

17. Hernandez-Abrego, G., and Marino, J.B. Contextual confidence measures for continuous speech recognition. In H. Abut and L. Onural (eds.), *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3. Los Alamitos, CA: IEEE Computer Society Press, 2000, pp. 1803–1806.

18. Higgins, E.L., and Raskind, M.H. Speaking to read: The effects of continuous vs. discrete speech recognition systems on the reading and spelling of children with learning disabili-

ties. *Journal of Special Education Technology*, *15*, 1 (Winter 2000) (available at jset.unlv.edu/15.1/higgins/first.html).

19. Hoffman, T. Speech recognition powers utility's customer service. *ComputerWorld*, September 12, 2005 (available at www.computerworld.com/managementtopics/management/helpdesk/story/0,10801,104535,00.html).

20. Kemp, T., and Schaaf, T. Estimating confidence using word lattices. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (eds.), *Proceedings of the Fifth European Conference on Speech Communication and Technology.* Bonn, Germany: International Speech Communication Association, 1997, pp. 827–830.

21. Krahmer, E.; Swerts, M.; Theune, M.; and Weegels, M. Error detection in spoken human–machine interaction. *International Journal of Speech Technology*, *4*, 1 (March 2001), 19–30.

22. Lai, J., and Vergo, J. MedSpeak: Report creation with continuous speech recognition. In S. Pemberton (ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM Press, 1997, pp. 431–438.

23. Levine, H.G., and Rossmoore, D. Diagnosing the human threats to information technology implementation: A missing factor in systems analysis illustrated in a case study. *Journal of Management Information Systems*, *10*, 2 (Fall 1993), 55–74.

24. Liu, Y. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. dissertation, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 2004.

25. Lubert, J.; Kotler, A.; Shein, F.; and Tam, C. Speech recognition. SNOW, Toronto, ON, 1998 (available at snow.utoronto.ca/best/special/speechrecognition.html).

26. Maison, B., and Gopinath, R. Robust confidence annotation and rejection for continuous speech recognition. In V.J. Mathews and A. Swindlehurst (eds.), *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2001, pp. 389–392.

27. Mangu, L., and Padmanabhan, M. Error corrective mechanisms for speech recognition. In V.J. Mathews and A. Swindlehurst (eds.), *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2001, pp. 29–32.

28. Mann, W.C., and Thompson, S.A. Rhetorical structure theory: A theory of text organization. In L. Polanyi (ed.), *The Structure of Discourse.* Norwood, NJ: Ablex, 1987, pp. 85–96.

29. Mao, J.-Y., and Benbasat, I. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, *17*, 2 (Fall 2000), 153–180.

30. McTear, M.F. Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, *34*, 1 (March 2002), 90–169.

31. Nunamaker, J.F., Jr.; Konsynski, B.R.; Chen, M.; Vinze, A.S.; King, D.R.; and Heltne, M.M. Knowledge-based systems support for information centers. *Journal of Management Information Systems*, *5*, 1 (Summer 1988), 6–24.

32. Pao, C.; Schmid, P.; and Glass, J. Confidence scoring for speech understanding systems. In R.H. Mannell and J. Robert-Ribes (eds.)*, Proceedings of the Fifth International Conference on Spoken Language Processing.* Canberra: Australian Speech Science and Technology Association, 1998, pp. 815–818.

33. Pradhan, S.S., and Ward, W.H. Estimating semantic confidence for spoken dialogue systems. In F.J. Taylor, J. Principe, and H. Bourlard (eds.), *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2002, pp. 233–236.

34. Ringger, E.K., and Allen, J.F. Error correction via a post-processor for continuous speech recognition. In M.H. Hayes and M.A. Clements (eds.), *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 1996, pp. 427–430.

35. Robertson, J.; Wong, W.Y.; Chung, C.; and Kim, D.K. Automatic speech recognition for generalised time based media retrieval and indexing. In W. Effelsberg and B.C. Smith (eds.), *Proceedings of the Sixth ACM International Conference on Multimedia.* New York: ACM Press, 1998, pp. 241–246.

36. San-Segundo, R.; Pellom, B.; Hacioglu, K.; Ward, W.; and Pardo, J.M. Confidence measures for spoken dialogue systems. In V.J. Mathews and A. Swindlehurst (eds.), *2001 IEEE*

International Conference on Acoustics, Speech, and Signal Processing, vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2001, pp. 393–396.

37. Sarikaya, R.; Gao, Y.; and Picheny, M. Word level confidence measurement using semantic features. In W. Siu, A.G. Constantinides, and Y. Chan (eds.), *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1. Los Alamitos, CA: IEEE Computer Society Press, 2003, pp. 604–607.

38. Sarma, A., and Palmer, D.D. Context-based speech recognition error detection and correction. In J.B. Hirschberg, S. Dumais, D. Marcu, and S. Roukos (eds.), *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics 2004: Short Papers.* East Stroudsburg, PA: Association for Computational Linguistics, 2004, pp. 85–88.

39. Sears, A.; Feng, J.; Oseitutu, K.; and Karat, C.-M. Hands-free speech-based navigation during dictation: Difficulties, consequences, and solutions. *Human–Computer Interaction*, *18*, 3 (2003), 229–257.

40. Sears, A.; Karat, C.-M.; Oseitutu, K.; Karimullah, A.; and Feng, J. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society*, *1*, 1 (June 2001), 4–15.

41. Skantze, G., and Edlund, J. Early error detection on word level. In B. Milner (ed.), *Proceedings of COST278 and ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction.* Bonn, Germany: International Speech Communication Association, 2004 (available at www.isca-speech.org/archive/robust2004/rob4_17.html).

42. Suhm, B.; Myers, B.; and Waibel, A. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer–Human Interaction*, *8*, 1 (March 2001), 60–98.

43. Weintraub, M.; Beaufays, F.; Rivlin, Z.; Konig, Y.; and Stolcke, A. Neural-network based measures of confidence for word recognition. In M.K. Lang and H. Hoge (eds.), *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2. Los Alamitos, CA: IEEE Computer Society Press, 1997, pp. 887–890.

44. Wendemuth, A.; Rose, G.; and Dolfing, J.G.A. Advances in confidence measures for large vocabulary. In D. Cochran and A. Spanias (eds.), *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2. Los Alamitos, CA: IEEE Computer Society Press, 1999, pp. 705–708.

45. Wessel, F.; Schluter, R.; and Ney, H. Using posterior probabilities for improved speech recognition. In H. Abut and L. Onural (eds.), *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3. Los Alamitos, CA: IEEE Computer Society Press, 2000, pp. 1587–1590.

46. Wessel, F.; Schluter, R.; Macherey, K.; and Ney, H. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, *9,* 3 (March 2001), 288–298.

47. Young, S.R. Detecting misrecognitions and out-of-vocabulary words. In *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2. Los Alamitos, CA: IEEE Computer Society Press, 1994, pp. 21–24.

48. Zhang, D., and Adipat, B. Challenges, methodologies, and issues in the usability testing of mobile applications. *International Journal of Human–Computer Interaction*, *18*, 3 (July 2005), 293–308.

49. Zhang, R., and Rudnicky, A.I. Word level confidence annotation using combinations of features. In P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan (eds.), *Proceedings of the Seventh European Conference on Speech Communication and Technology.* Bonn, Germany: International Speech Communication Association, 2001, pp. 2105–2108.

50. Zhou, L.; Shi, Y.; Feng, J.; and Sears, A. Data mining for detecting errors in dictation speech recognition. *IEEE Transactions on Speech and Audio Processing*, *13*, 5 (September 2005), 681–688.

51. Zhou, Z., and Meng, H. A two-level schema for detecting recognition errors. In S.H. Kim, S. Lee, Y. Oh, and Y. Lee (eds.), *Proceedings of the Eighth International Conference on Spoken Language Processing.* Bonn, Germany: International Speech Communication Association, 2004, pp. 449–452.

Appendix. A Taxonomy of Cues to Error Detection (CERD)

Linguistics-Based (L)

- Phonological (P)
  - Disfluency (LP.1)
    - False start (LP.1.1)
    - Repetition (LP.1.2)
  - Phonetic similarity (LP.2)
  - Word split (LP.3)
- Morphological (M)
  - Part-of-speech confusion (LM.1)
    - Conjunction (LM.1.1)
  - Open-class word choice (LM.2)
  - Closed-class word choice (LM.3)
    - Preposition (LM.3.1)
    - Determiner (LM.3.2)
  - Nonword choice (LM.4)
- Syntactic (G)
  - Phrase structure (LG.1)
    - Phrase head (LG.1.1)
    - Parallel structure (LG.1.2)
    - Incomplete phrase (LG.1.3)
    - Modifier (LG.1.4)
    - Common phrase (LG.1.5)
  - Sentence structure (LG.2)
    - Beginning/ending (LG.2.1)
    - Tense/voice/aspect/modal (LG.2.2)
    - Number disagreement (LG.2.3)
    - Dangling words (LG.2.4)
    - Incomplete sentence (LG.2.5)
    - Subject/object (LG.2.6)
    - Person disagreement (LG.2.7)
- Semantic (S)
  - Making no sense (LS.1)
  - Incompatible semantics (LS.2)
    - Subject–object (LS.2.1)
    - Preposition–object (LS.2.2)
    - Modifier–head (LS.2.3)
    - Subject–predicate (LS.2.4)
    - Predicate–object (LS.2.5)
    - Two constituents (LS.2.6)
- Discourse (D)
  - Out of context (LD.1)

    - Preceding (LD.1.1)
    - Following (LD.1.2)
  - Coreference (LD.2)
  - Contradictory information (LD.3)
  - Background scenario (LD.4)

Hypotheses-Based (H)

- Word (W)
  - Word hypotheses (HW.1)
  - Confidence scores (HW.2)
  - Highest confidence score (HW.3)
- Utterance (U)
  - Utterance hypotheses (HU.1)
  - Utterance length (HU.2)
  - Path ratio (HU.3)

Others (O)

- Error (E)
  - Preceding error (OE.1)
  - Following error (OE.2)
- Style (S)
  - Contraction (OS.1)
  - Double negation (OS.2)
  - Letter case (OS.3)
  - Misplacement (OS.4)
- Repetition (R)
  - Redundancy (OR.1)