Modeling the Speed of Text Entry with a Word Prediction Interface

Heidi Horstmann Koester and Simon P. Levine

***Abstract*—This study analyzes user performance of text entry tasks with word prediction by applying modeling techniques developed in the field of human-computer interaction. Fourteen subjects transcribed text with and without a word prediction feature for seven test sessions. Eight subjects were able-bodied and used mouthstick typing, while six subjects had high-level spinal cord injuries and used their usual method of keyboard access. Use of word prediction decreased text generation rate for the spinal cord injured subjects and only modestly enhanced it for the able-bodied subjects. This suggests that the cognitive cost of using word prediction had a major impact on the performance of these subjects. Performance was analyzed in more detail by deriving subjects' times for keypress and list search actions during word prediction use. All subjects had slower keypress times during word prediction use as compared to letters-only typing, and spinal cord injured subjects had much slower list search times than able-bodied subjects. These parameter values were used in a two-parameter model to simulate subjects' word entry times during word prediction use, with an average model error of 16%. These simulation results are an encouraging first step toward demonstrating the ability of analytical models to represent user performance with word prediction.**

I. Background

COMPUTER-BASED augmentative and alternative communication (AAC) systems provide people who have severe disabilities with the opportunity to communicate independently in the areas of speech, writing, and computer applications. A major goal in the design and prescription of these systems is to provide the user with the fastest means of communication possible. A variety of techniques designed to enhance user performance are currently used in AAC systems, including word abbreviations [1], [2], message encoding [3], [4], and word prediction [5], [6]. There continues to be a need for greater understanding of the efficacy of these systems.

A primary aim in most rate enhancement approaches is to reduce the motor requirements placed on the user. This is clearly an important goal, since the vast majority of users have severe physical impairments. However, a frequent consequence of reducing motor requirements is to increase the cognitive and perceptual loads on the user [4], [7], [8]. The net balance of this trade-off determines whether the user's overall performance will be enhanced or inhibited with a system [9].

Manuscript received November 17, 1993; revised August 22, 1994. This work was supported in part by the National Science Foundation, in part by the University of Michigan Rackham School of Graduate Studies, and in part by the University of Michigan Rehabilitation Engineering Program.

The authors are with the Rehabilitation Engineering Program, Graduate Bioengineering Program, Department of Physical Medicine and Rehabilitation, University of Michigan, 1C335 University Hospital, Ann Arbor, MI 48109-0032 USA.

IEEE Log Number 9405747.

This paper focuses on user performance with word prediction systems in particular and how it is affected by the trade-off between decreased motor and increased cognitive loads. Word prediction systems attempt to predict the word intended by the user by presenting the user with a set of word choices. Word prediction choices are typically displayed in a short list and are refined as the user selects additional letters. Since many words can be completed by choosing from the list rather than through letter-by-letter spelling, the number of selections required per word can be substantially reduced. Keystroke savings provided by several commercial word prediction systems have been measured in the range of 37–47% [10], with clinical reports ranging from 23–58% [5], [11]-[13].[1]

Keystroke savings represents the extent to which word prediction reduces the motor requirements on the user relative to letter-by-letter spelling. This benefit comes at the cost of additional cognitive and perceptual activities required to use the system. These include the visual search of the word list and the subsequent decision about whether the list contains the desired word. An additional source of cognitive load may be the processing involved in planning use strategies (e.g., deciding when to search) and guiding overall activity [14]–[16].

Evidence that these additional cognitive loads can have a negative effect on user performance is shown in Fig. 1. The figure shows the improvements in text generation rate with word prediction as reported in the literature for 13 individuals, relative to the keystroke savings achieved by these individuals [5], [11], [12], [17], [18]. It also shows what the rate improvements *would be* if there were no time cost due to additional cognitive and perceptual activities [9]. All but two of these individuals achieved less than this ideal improvement, which provides indirect yet strong evidence that the additional cognitive and perceptual activities reduce the benefit of decreased motor requirements. More direct evidence comes from our recent study on able-bodied users of scanning systems, in which use of word prediction slowed the rate of selecting items (i.e., letters and/or words) by 30–40% compared to letters-only typing [17].

In addition to providing evidence of cognitive cost, these data also show a large diversity in the effect of word prediction on text generation rate. This diversity may be partially due to differences in methodologies between studies, but it also suggests that the effect of word prediction depends on the

[1] Keystroke saving is measured as 1-(keystrokes required/ characters generated). Keystrokes are broadly defined to include keypresses in a direct selection system, as well as items selected in other ways, such as through scanning or Morse code.

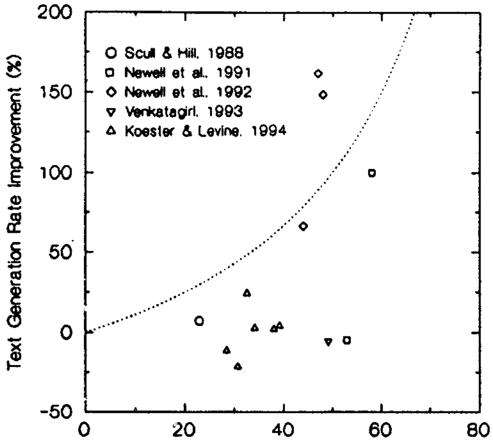

Fig. 1. Reported improvements in text generation rate with word prediction as a function of keystroke savings. Each point corresponds to the performance of a single individual. The dotted line shows what rate improvement *would be* if there were no cognitive time costs associated with use of word prediction.

specific characteristics of the individual user and the context of use. For some individuals, under some conditions, the benefit of keystroke savings seems to outweigh the cost of additional cognitive activities, resulting in a welcome improvement in overall speed of text generation, while for other individuals, the opposite is true. A goal of our research is ultimately to determine the conditions under which word prediction improves text generation rate and those under which it does not.

An empirical approach to pursuing this goal involves measuring the performance of a variety of users under a range of conditions and attempting to deduce the underlying principles of user performance from the resulting data. There is a great need for further empirical information gathered under well-defined conditions, particularly in light of the relatively sparse and diverse data reported to date, as seen in Fig. 1. However, a limitation of an empirical focus is that only a small subset of user-system combinations can be studied, so it is difficult to extrapolate to conditions that have not been empirically examined.

One way to move beyond the limitations of a purely empirical approach is through the development of analytical models that integrate information about the system and user to predict the user's performance [13], [14], [19], [20]. These build on empirical data and support the simulation of user performance across a range of conditions. Accurate analytical models could provide AAC system developers with a means of evaluating the consequences of design decisions, to support the development of an optimal design. For clinicians, models could aid system prescription and configuration by estimating user performance with a range of candidate systems.

The prognosis for user performance modeling in AAC is somewhat controversial. Questions exist about whether accurate models of user performance can be developed, due to variation in the abilities of the user population and the different approaches users may take toward a particular system [12]. The potential benefits of the modeling approach have also been recognized, and several research efforts have been aimed at quantitative model development [13], [19], [20]–[23]. Many of the models developed to date evaluate systems based primarily on motor efficiency [20], [22], [23], so they are not well-suited to represent systems like word prediction, in which cognition and perception have an important impact. Models which have explicitly included cognitive and perceptual processes have provided important conceptual frameworks, but very little work has been done to compare their quantitative predictions to actual user performance [13], [19], [21]. These limitations in model structure and/or the extent of model validation have hampered the success of previous efforts, so the question of whether an accurate model of user performance can be developed in AAC remains open.

Outside of AAC, a great deal of research attention has focused on the development of user interface modeling techniques [15], [16], [24]. The research presented here is based on one such technique, called the Keystroke Level Model (KLM) [15]. This technique provides a means of identifying the cognitive, perceptual, and motor activities that a user must perform for a particular task. The time required for executing that task is then predicted by summing the times for each component activity. In the case of text entry with a word prediction system, the unit task is the entry of a single word, accomplished through a series of letter and word list selections, each of which involves cognitive, perceptual, and motor component actions.

A primary feature of the KLM is its ability to accurately account for user performance using only a small number of parameters. For example, [16] was able to predict the time to enter spreadsheet commands with a 26% error, using only two user parameters, one for keypress time and one for general mental time. The use of the aggregate mental operator illustrates the emphasis of the model on providing useful approximations to cognitive costs, rather than precise psychological models for each cognitive process. It is recognized that this feature is a potential source of model error, but it is important to determine whether useful accuracy can be obtained despite the simplifying assumptions used. Knowledge of the limitations of these assumptions provide guidance for subsequent revisions of the model if necessary. The flexibility in the KLM technique, as well as its proven accuracy in modeling similar tasks, were the reasons for choosing the KLM as the basis for this work.

II. Purpose

This research is part of a long-term program to gain greater empirical understanding of user performance with AAC systems and to develop analytical models that can accurately simulate expected performance. The current study focuses on word prediction and addresses the following specific issues:

1) Available data on user performance with word prediction suggests that the time required for additional cognitive and perceptual processes involved in the use of word prediction will at least partially offset the benefit of decreased motor requirements. This study is intended to extend the available data base by employing both able-bodied and physically disabled subjects across a multi-

session protocol. We hypothesize that text generation rates for all subjects with word prediction will be lower than those expected based solely on consideration of keystroke savings. Further, we propose to examine the source of the cognitive loads in greater detail than has been reported previously, through the derivation of subjects' list search and keypress parameter times.

2) The potential for analytical user performance models in AAC has not yet been realized, and it is not clear if this is due to limitations in previous modeling work or to more serious theoretical problems. A main goal of this study is to address this controversy. We hypothesize that a two-parameter model of user performance can be developed using KLM techniques which will simulate word entry time with an accuracy at least as good as that reported for other applications of the KLM (below 25–30% error). In this initial effort, the two parameter values for keypress and list search times will be derived for individual subjects based on their performance data.
3) Within the general issue of modeling feasibility, a main question is how well a model can accommodate differences between individual users or groups of users. This study addresses part of this broad issue by comparing model accuracy for able-bodied subjects to a group of spinal cord injured subjects. We hypothesize that model simulations will be equally accurate for able-bodied and physically disabled subjects. Any differences in observed performance between these two groups will be accounted for in the model by using different user parameter values within the same two-parameter model structure [25]. We expect the major difference in user parameter values to be in those that represent motor activities, rather than cognitive activities, since these subjects will have only physical disabilities.
4) A fourth goal is to explore the potential of using an analytic model to identify optimal strategies for system use. As a first step toward this long-term goal, we hypothesize that the two-parameter model will be equally successful in simulating performance under different strategies of use. Additionally, we expect user parameter values to be independent of strategy used, even if overall performance is not, since the parameters are intended to represent fairly low-level building blocks of overall performance.
5) Finally, the study addresses the accuracy of model simulations across a range of usage conditions, with the hypothesis that model accuracy will not change as subjects gain experience with word prediction during the experiment or as the keystroke savings of the system is varied. As in #3 above, any differences in performance with practice will be accounted for by different user parameter values within the same two-parameter model structure.

To test these hypotheses, an experiment was conducted to measure user performance with and without word prediction, as a source of model validation data as well as a contribution to empirical understanding. User performance was modeled using KLM techniques with parameter values derived from the subject data. Actual subject performance was then compared to modeled performance. The modeling results reported here represent one step toward a thorough assessment of the model's accuracy, as a necessary prerequisite to applying the model in clinical and design situations.

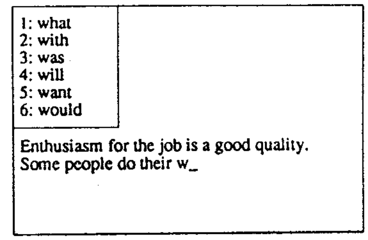

Fig. 2. Schematic representation of the Letters+WP system display. The six-word list is fixed in the upper left corner, with the transcribed text displayed below the list. A sample of actual text transcribed by subjects is shown, and the list contents are those that follow selection of the letter w.

III. Methods

A. Subjects

Fourteen subjects were employed. All subjects shared the following characteristics: at least some post-secondary education; frequent computer use and high familiarity with the standard keyboard layout; no significant prior experience with word prediction; and no reported cognitive, perceptual, or linguistic disabilities. Eight of the subjects were able-bodied, and the remaining six had spinal cord injuries at levels ranging from C4–C6.

B. Interfaces

The two interfaces used in the study were developed by the investigators specifically for research purposes, to provide sufficient control over the means of data collection. Both interfaces used direct selection on the standard computer keyboard as the basic input method. Able-bodied subjects used mouthstick typing to access the keyboard, while subjects with spinal cord injuries used their usual method of keyboard access, which was mouthstick typing for two of the subjects and hand splint typing for the other four. The first interface, referred to as "Letters-only," simply involved letter-by-letter spelling. The second interface, referred to as "Letters+WP," used single letter entry augmented by a word prediction feature. A six-word prediction list with a fixed word order was used and presented vertically in the top left corner of the screen, as shown in Fig. 2.

C. Experimental Design

The protocol involved a three-session training phase and a seven-session testing phase. The testing phase employed an alternating treatments design, in which subjects' text transcription performance with and without word prediction was recorded in each test session. The keystroke savings provided by word prediction was fixed across Sessions 1–4 and varied in Sessions 5–7 (as discussed in more detail below). Each subject was randomly assigned to one of two different strategies with

TABLE I
THE FOUR SUBJECT GROUPS

	SCI No	SCI Yes
Strategy 1	AB1 (n=4)	SCI1 (n=3)
Strategy 2	AB2 (n=4)	SCI2 (n=3)

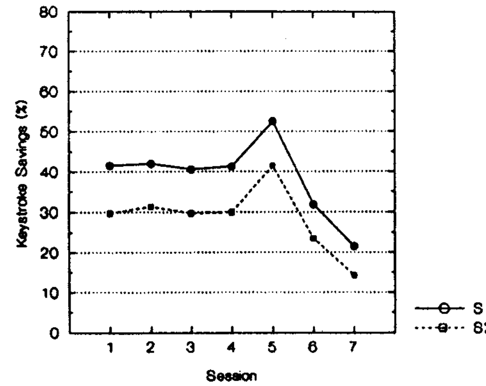

Fig. 3. Keystroke savings provided by the Letters+WP system at each session for each strategy.

which they were to use the word prediction feature. These strategies are discussed in more detail below. The assignment of subjects to the four groups is shown in Table I.

D. Procedures

Subjects were tested in individual sessions which were conducted in laboratory space for eleven subjects and at subjects' homes for the three spinal cord injured subjects who had difficulty arranging travel to the university. Subjects took an average of 21 days to complete the protocol.

Training began with two sessions of practice using the Letters-only system. Each able-bodied subject was provided with a 17″ anodized aluminum mouthstick to use for the duration of the study,[2] while the spinal cord injured subjects used their own mouthstick or typing splints. For able-bodied subjects, the keyboard was placed at standard desk height and tilted at an angle of 45 degrees relative to the desk surface. For spinal cord injured subjects, the keyboard was placed to match their normal set-up; all used a flat keyboard. In the first training session, subjects were instructed in the transcription task and proper use of the mouthstick was demonstrated for able-bodied subjects. Subjects were given the goal of typing as quickly as possible, while keeping mistakes to a minimum. They then practiced for six blocks of text (four sentences each) over two sessions. After each block of text, subjects were asked to rate the difficulty of the task on a continuous scale ranging from "Very Easy" to "Very Difficult."

The third training session introduced subjects to the word prediction feature and their assigned strategy for its use. The rules for the two strategies were defined as follows:

Strategy 1. Search the list before every selection.

Strategy 2. Choose the first two letters of a word without searching the list, then search the list before each subsequent selection.

For both strategies, an exception to these rules occurred when the word list was empty, in which case a list search was not required. These strategies were chosen to be realistic enough to represent at least a subset of actual user approaches, simple enough to be learned in a single training session, and distinct enough to yield measurable performance differences. Subjects were asked to follow the rules as closely as possible. All subjects practiced using their strategy for four blocks of text (4 sentences each), which was sufficient for each to use the strategy correctly without prompting.

Each of the seven test sessions involved four sentences of warm-up using word prediction, an eight-sentence test with word prediction, then a two-sentence typing test. Text blocks were drawn from published typing tests, matched with respect to syllable intensity, average word length, and percent of words that occur with high frequency [26]. The texts were carefully revised to provide the same level of keystroke savings across Sessions 1–4 and systematic variation of keystroke savings in Sessions 5–7. Fig. 3 shows the specific keystroke savings across sessions for each strategy of word prediction use. (Note that the text transcribed was identical for both strategies; the difference in keystroke savings was caused only by differences in the strategies.)

Sentences were presented singly on index cards. Subjects were given twenty seconds to read the sentence before an audio cue signalled them to begin transcription. Errors could be corrected by selecting the "Backspace" key as well as a special key for correcting word list selections. The sentence card remained in view for reference throughout transcription.

E. Data Collection

All items selected by subjects were timed and stored by the software in real time. Entries were also encoded to store various information such as the type of selection made (i.e., a single letter or a word list selection) and the number of words in the list when the item was selected. The raw data was used to produce an entry log, in which each line shows the selected item, its various characteristics, and the time at which it was selected.

All sessions were videotaped, with the camera focused on the subject's face, close enough to determine easily the direction of eye gaze. Keypresses were recorded on the video using a mirror, placed behind the subject to reflect a view of the keyboard into the camera, and a speech synthesizer, which echoed the selected item onto the audio track (without being audible to the subject). The camera's clock was synchronized with that of the computer, so the times on the videotape matched those on the entry log.

An experimenter was present throughout each session to record observations of subject behavior. In addition to the difficulty rating described above, subject comments were solicited after each session. Subjects were also given immediate feedback on their text generation rate with each system.

[2] AdLib Incorporated, 5142 Bolsa Avenue, Suite 106, Huntington Beach, CA 92649.

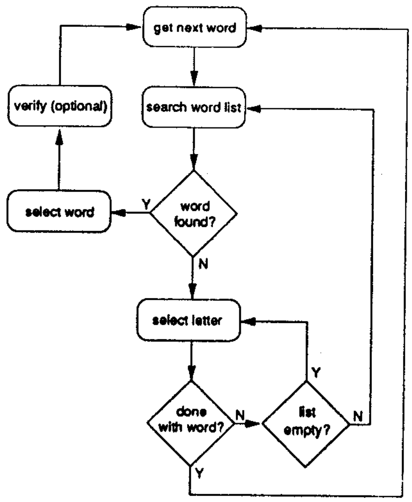

Fig. 4. Flow chart representing activities required during use of Strategy 1 with the Letters+WP system.

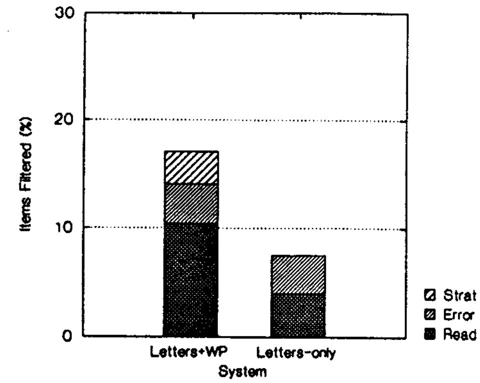

Fig. 5. Amount of data filtered for each system as a percent of total number of items selected, averaged across subjects and sessions. Also shown is the relative contribution of the three filtering categories: "Strat" for items not consistent with the assigned strategy, "Error" for erroneous selections, and "Read" for items following subjects' referral to the text card.

F. Data Filtering

The raw data were filtered to remove events judged to be in any of the following three categories. The first was a general category of text errors and error corrections, including typographical as well as transcription errors. The second included all words that were not entered in a manner consistent with the assigned strategy. Events in these two categories were identified by comparing the subject's generated text to an error-free template. The final category consisted of "card reads," or times when the subject referred back to the text card during transcription, as identified through analysis of the videotape records. Carriage returns and periods were filtered out as well.

The process of identifying and coding events to be filtered was performed by the first author and a trained assistant. Interrater reliability was measured at 99.4%, based on a sample of four sessions analyzed by both raters. A point-to-point reliability measure was used, in which the total number of agreements between raters is divided by the total number of agreements and disagreements [27].

G. Dependent Measures of User Performance

Text generation rates for the Letters+WP and Letters-only systems were measured for each subject at each test session by dividing the number of characters generated during the test by the total time required to generate those characters. Items that were filtered out were not counted either in the number of characters generated nor in the total time.

H. Measurement of User Parameters:

The first step in measuring user parameters was to determine the parameters most important to task execution time. This was done by analyzing the task of entering words for each of the word prediction strategies. As an example, the flow chart of hypothesized user activity for Strategy 1 is shown in Fig. 4. The major activities for both strategies are keypresses (to select a letter or a word) and list searches, so these were chosen as the two user model parameters. While additional parameters could have been defined, only two were used, in order to test the accuracy of a more parsimonious model. A consequence of this choice is that although each parameter primarily represents the activity for which it is named, it may include other components as well. For example, the keypress parameter reflects the motor component of selecting an item, but it may also incorporate more subtle activities such as verifying accuracy, retrieving from memory the next word to be typed, or retrieving the rule that guides the next action to be made. Similarly, the list search parameter may also include these extra activities, in addition to the list search itself. Fortunately, such integration is consistent with the spirit of previous KLM studies, in which a single mental time parameter has often been used for all cognitive actions [15].

Durations for the component actions of list search and keypress while using Letters+WP were derived from the filtered data for each subject. The technique used for this followed the subtractive methods of [15], [16]. Based on the strategy used with Letters+WP, each selection was labelled according to whether it involved a keypress preceded by a list search or a keypress with no list search. For example, when using Strategy 2, the first two letters of every word involved no list searches, so they were labelled as keypress-only. The third letter, however, did include a list search, so it was labelled as a list search-plus-keypress. The keypress time (t_k) during use of Letters+WP was then calculated by averaging the times for all keypress-only selections in the session. The list search time (t_s) was derived by subtracting one t_k from the time recorded for each list search-plus-keypress selection, then averaging the remaining times. In all, 98 pairs of parameter values were derived in this way (14 subjects $\times$ 7 sessions).

I. Model Simulations

Using these parameter values, simulations of the time to enter each word during use of Letters+WP were performed as follows. A model value for each item selection was calculated based on whether that selection involved a keypress only (t_k) or a list search-plus-keypress ($t_s + t_k$). The values were summed for each item in a word to yield an entry time for