Modeling and analysis of dyslexic writing using speech and other modalities

Jerome Elkind and Jeff Shrager

Introduction

In Chapter 5 Newell et al. discuss communication and conversation aids for people with speech and language impairments. They review the problems that these people have producing natural sounding speech at acceptable rates and with low response latency, and they discuss a number of computer based prostheses to assist them. In this chapter we focus on the specific problem that people with dyslexia have in communicating through the written medium, and we analyze in detail the performance obtained from several ways of using computers to support writing. The objective is to present a general methodology for analyzing compensatory aids designed to help people with disabilities, to show how this methodology can be used to understand how writing speed is affected by the characteristics of the compensatory aids, in particular a speech recognition system, and to suggest ways in which the speech recognition system could be redesigned to improve performance.

Many designers of complex devices believe that better psychological and physiological models will lead to better design (e.g. Sheridan & Ferrell, 1974; Card, Moran, & Newell, 1983; Elkind, Card, Hochberg, & Huey, 1990; Rouse, 1991). This view has led to a fruitful interaction between these human sciences and design that goes by the name "human factors." Human factors research has influenced the design of many systems, and designers themselves often engage in human factors observation or thought experiments.

What psychology can offer design, beyond what designers can gather on their own, is a set of models that are rigorously grounded and relatively general. Unfortunately, rigor and generality are accomplished through experimental controls and statistical analysis and thus at the expense of richness and detail. The resulting models, paradoxically, describe the normal case even if there is not even a single exemplar of that case. Individual differences dissolve into variance from the mean, or worse, are considered outside of the scope of the model. Thus the capabilities of people with impairments lie outside of general models almost
by definition, and the handicaps they experience are exacerbated by norm-based human factors designs.

One practical remedy for this state of affairs is to design systems in accordance with models that specify dimensions and ranges rather than particular values. For instance, one can imagine an automobile dashboard whose illumination color is adjustable to accommodate various sorts of color blindness. Such flexibility is not usually so straightforward to implement, and most human impairments do not lie along such simple dimensions. Moreover, the extreme case, even on simple dimensions, is often qualitatively different from the other ranges; it is a significant oversimplification, for example, to equate deafness with lesser hearing impairments.

A general model nonetheless has heuristic value as long as we are willing to modify it in significant ways when it fails to describe a specific person's condition adequately—thus bending the model to the person rather than the person to the model. The resulting new model can then usefully be studied to produce design proposals that are likely to address the specific needs of persons with impairments.

In this chapter we apply this method in the case of computer supported composition by dyslexic individuals. Dyslexia is a complex class of cognitive/linguistic impairments, principally of the ability to process written language. It is important to keep in mind that each dyslexic person has specific impairments from this complex class and must be understood individually. Assistive devices and remediation programs must be adjusted for individual differences.

We examine how dyslexia affects the writing process and how to improve the performance of dyslexics in computer supported writing and other text related tasks. We have studied transcription by dyslexic and nondyslexic persons in several settings: writing longhand without computer assistance, using a computer text editor, using a speech recognition prosthesis, and dictating to a typist. Here we develop a model of transcription activity and use the model to isolate points of difficulty for a dyslexic person. This analysis enables us to see clearly why the speech recognition prosthesis is less effective than expected in compensating for the dyslexic person's disability. Finally, from these analyses and from observations of the collaborative setting in which the writer is working with a human typist, we produce several suggestions to improve the effectiveness of speech recognition as a prosthesis for dyslexic writing.

Dyslexia

Dyslexia affects a large number of people. Exactly how large is difficult to pin down (Elkind, 1990). U.S. census data indicate that about 5% of public school students are enrolled in special education programs for the learning disabled, a disabilities category that includes dyslexia (U.S. Department of Commerce, 1987, 1989). Since not all students with learning disabilities are enrolled in these
programs, the fraction of students with this condition is probably higher. Speier and O'Connell (1990) cite data from the National Institutes of Health showing that 10% to 15% of the population have learning disabilities. Other data suggests that it is of the order of 7.5% of the population (Gaddes, 1985). Of these, somewhere between 60% (Gaddes, 1985) and 80% (Speier & O'Connell, 1990) would be considered dyslexic. Thus more than 10 million people, and perhaps as many as 20 to 30 million people, in the USA have this disability. Dyslexia can be a serious handicap that often blocks intelligent people from education and employment. Learning disability and dyslexia are associated with a large fraction of school dropouts and juvenile delinquents (Speier & O'Connell, 1990).

Dyslexia is a language disorder, particularly a disorder of written language. Dyslexics are slow to develop fluency in reading and writing and to automate the processes involved in these skills. Many dyslexics never become fluent, and their reading and writing remain inaccurate, slow, and limited in vocabulary. Dyslexics have difficulty associating the written form of language with its oral form. Their ability to decode words, to spell, and to recall letter sequences of words and even images of individual letters is frequently poor. Their sight recognition of words is limited. As a result, their reading requires conscious effort, is error prone, and is associated with poor comprehension.

Writing, because it is generative, is even more severely affected. The demands of spelling associated with writing – retrieving the correct sequence of letters in words or computing them from phonological rules – impose a substantial cognitive load that often interferes with the demands of organizing and expressing ideas in well formed sentences and paragraphs. Dyslexics often have difficulty in choosing correct prepositions and word endings and suffixes, and in verb–subject agreement. They omit words from sentences and telescope ideas so that thoughts are incomplete and conjunctives and transitions are missing. These difficulties, coupled with their meager writing vocabulary, leads to written work that is a poor reflection of their intellectual abilities. Sometimes these difficulties spill over to oral language, but usually written language is more severely affected. Dyslexics therefore are usually better able to express their thoughts orally. The use of a speech recognition prosthesis would thus seem to be an attractive method for increasing their fluency in composition.

Dyslexia results from deficits in language processing, memory, perception, and discrimination occurring either separately or in many different combinations (Doehring, Trites, Patel, & Pliedorwicz, 1981; Levine, 1987). The language processing system may have difficulty appreciating the phonetic elements of language or the morphology of language (e.g., prefixes, suffixes). It may have trouble associating words with meanings. There may be weakness in visual–aural associative memory, in sequential memory (grasping and retrieving sequences of letters, words, sounds), difficulty in storing and retrieving linguistically encoded data, difficulty in retrieving motor sequences for letter or word production.

Working memory deficits may make it difficult to recall the beginning of a word or sentence while reading the end of it or to retain the components of a sentence
or word while carrying out higher level processing. Sensory memory limitations often make it difficult for a dyslexic to capture visual or aural sequences. Recent results suggest that at least some dyslexics may have a dysfunction in the fast response path (magnocellular system) of the auditory and visual perceptual systems (Livingstone, Rosen, Drislane, & Galaburda, 1991). This complex of deficits affects writing as well as reading because the limitations make it difficult for a dyslexic writer to retrieve or compute the spelling of words, to use correct forms of words, and to find appropriate words to use.

**Comparative performance**

Our empirical and analytic methodology is deployed in two phases. In the first phase we observe dyslexic and nondyslexic individuals in typical writing activities. (We will use the name K for the dyslexic subject of the studies reported here, and S for the nondyslexic subject.) K is a 23-year-old male college sophomore. His reading is slow, and he has difficulty decoding and remembering written forms of words. He has great difficulty with spelling, limited immediate recall spelling vocabulary, difficulty in generating correct sentence syntax, and some word retrieval difficulty. He is very slow in producing written material. He is taking two-thirds of a normal course load at school and relies extensively on taped books for reading and on dictation for writing. The nondyslexic subject, S, is a 25-year-old female graduate student who has excellent reading and writing skills and speed.

We investigated three kinds of activities:

1. transcription, in which the individual has to copy a document;
2. performance on a battery of diagnostic tests; and
3. composition of an essay describing a picture.

Four output methods were studied:

1. longhand writing;
2. into a computer word processor (Microsoft Word running on a Macintosh SE computer);
3. dictating to a computer based speech recognizer (*DragonDictate*); \(^1\) and

---

\(^1\) *DragonDictate™* is a product of Dragon Systems Inc., Newton MA. We used version 1.0 of the DragonDictate software. It had a 30,000-word vocabulary that adapted to the user's speech patterns and vocabulary as it was being used. It is not a continuous speech recognizer but requires words to be separated by a brief pause. The user must check that each word is recognized correctly. The recognizer appends its first-choice word to the line of text shown on the computer screen and then posts a short ordered list of other candidate words on the screen. The user can select a word from this list by saying, for example, “Choose 3.” If the word spoken is not on the list, the user can
Modeling and analysis of dyslexic writing

Table 7.1. Typical rates for producing text by a dyslexic and nondyslexic in three activities using four methods for capturing text

<table>
<thead>
<tr>
<th>Activity</th>
<th>Method</th>
<th>Writing</th>
<th>Typing</th>
<th>Speech recognizer</th>
<th>Collaborative transcriptiona</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(words/min)</td>
<td>(words/min)</td>
<td>(words/min)</td>
<td></td>
</tr>
<tr>
<td>Transcription</td>
<td>Dyslexic</td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Nondyslexic</td>
<td>27</td>
<td>47</td>
<td>NA</td>
<td>155</td>
</tr>
<tr>
<td>Diagnostic test</td>
<td>(simple sentences)</td>
<td>Dyslexic</td>
<td>14</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Nondyslexic</td>
<td>25</td>
<td>35</td>
<td>NA</td>
<td>68</td>
</tr>
<tr>
<td>Essays</td>
<td>Dyslexic</td>
<td>5</td>
<td>6</td>
<td>NA</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Nondyslexic</td>
<td>33</td>
<td>NA</td>
<td></td>
<td>49</td>
</tr>
</tbody>
</table>

a Excludes typing time.

4. collaborative transcription, where the writer dictates to a human transcriber who types into the word processor.

Each activity was videotaped. From analysis of the videotapes we obtained measures of overall performance that enable us to compare the several activities and methods. Table 7.1 shows the text production rates of the two subjects in the three activities using the four output methods. First consider the performance of K, the dyslexic subject, in the transcription task. Handwriting, typing, and the speech recognizer gave roughly equivalent text generation rates, a slow 9–12 words/minute. For this individual, the speech recognizer was not faster than writing or typing; in fact it was a bit slower, for reasons we discuss below. However, when dictating to a human transcriber, the subject was able to perform at a rate of 69 words/minute, six to eight times faster than with the other methods. This is approximately the oral reading rate for this subject. In computing the rate for dictating to a human transcriber, we subtracted from the total task completion time the amount of time the subject spent waiting for the typist to complete the typing of previously dictated material. Thus, the rate shown in Table 7.1 for the collaborative transcription output method is the rate that would have been achieved if the typist had been infinitely fast.

Subject S, the nondyslexic, had good typing skills (47 words/minute) and was almost two times faster typing than writing by hand (27 words/minute). When
dictating to a human transcriber, S could produce text at a rate of 155 words/minute (after subtracting time spent waiting for the typist to complete previously dictated text), more than three times faster than typing. S was two to four times faster than K, depending upon the method.

The diagnostic test included in Table 7.1 required the subject to look at sketches of simple scenes (for example, a boy holding a baseball bat and looking anxiously at a broken house window) and then rapidly to write short sentences describing the scenes.1 This is an example of a very simple composition task. In this activity the subject did not have to read text sentences as in the transcription task, but instead had to examine the sketch and generate a descriptive sentence. For K, the text production rates in this diagnostic test were similar to those for transcription (+/- 20%). Collaborative transcription remained six to eight times faster than the other output methods. For S, text production rates for this task were slower than for transcription. The decrease is especially large for collaborative transcription; text production rates declined by more than a factor of two from 155 words/minute to 68 words/minute. For S, this simple composition task apparently took more time than reading, whereas for K it took about the same time. K and S had roughly equivalent rates when dictating, whereas with the other methods K was slower by a factor of two to four.

In the essay composition activity the subjects looked at a sketch of a more complex scene (for example, a group of people walking and playing on a city street on a windy day) and were asked to write an essay of two to three paragraphs describing the scene.2 This is a more demanding composition activity, and it is not surprising to see the word rates for K decline substantially. His word production rates were about half those obtained in the other activities. Typing and speech recognizer rates were only 5–6 words/minute. Collaborative transcription was reduced to 29 words/minute, although it was still five to six times faster than the other methods. Working collaboratively with a human transcriber is clearly a much more effective way for K to write than any of the other methods.

Rates for S also declined, but by smaller amounts. There was a small reduction in typing rate to 33 words/minute and a greater reduction in dictation rate to 49 words/minute. Thus S does not gain not much advantage by working collaboratively with a human transcriber. Her typing speed is about as fast as her ability to compose text.

It is important to note the difference in word production rates between the dyslexic and the nondyslexic. When K had to generate typewritten text, his rate

---

1 Test material was taken from the Writing Fluency component of the Woodcock Johnson Tests of Achievement (WJ-R) (Woodcock & Johnson, 1989).

2 Test material was taken from the PIAT-R test of written expression (Markwardt, 1989).
was one third to one sixth that of S, depending upon the task. When dictating to a human transcriber, his rates were more competitive, from almost equal to about one half of S's rate. Clearly he is at a substantial disadvantage compared with his nondyslexic peers in school or in work activities that require the production of written material unless he can dictate. A computer based writing prosthesis that would allow him to use speech in a way that mimics his interaction with a human transcriber would be of enormous value.

The DragonDictate system did not provide him this facility. One reason is that the DragonDictate requires considerable attention to control, error detection, and error correction functions. Another reason may be that K was not really expert in the use of the DragonDictate. Although he had about 20 hours of use of the system prior to these tests and was reasonably proficient with it, more experience would probably have led to greater facility and better strategies for control and error handling. Although the DragonDictate was not faster than typing or writing, K reported lower cognitive stress when using it, since it eliminated most of the cognitive load associated with spelling.

Flow models of language interaction with computers

In the second phase of analysis, we produced detailed activity flow models of the interaction between computers and dyslexic and nondyslexic persons in performing the transcription tasks. These models identify the processes that take time in the activity. Individual processes might involve the person (whether in cognitive or physical action), the computer, some other aspect of the environment (e.g., the document being transcribed), or relationships between these.

We think of these models as part of the overall writing activity as studied by Hayes and Flower (1980). Figure 7.1 is an extension of the model postulated by Hayes and Flower. Their model, derived from detailed analysis of protocols obtained from subjects performing writing assignments, consists of three major processes: planning the composition, translating the material developed under the plans into written sentences, and reviewing the written text to improve its quality. These processes operate in a context defined by the writing assignment and the writer’s experience.

The Hayes and Flower model — and most other studies of writing — assumes an unimpaired writer. The model does not dwell on the processes for producing and capturing text in written form or for reading existing text. For most people, once a sentence is formulated, text production is a highly automated process; for the dyslexic it requires conscious and often great effort. To account for the difficulty associated with these activities, we have added to the original Hayes and Flower model components concerned with the production and capture of text. The text production process converts the words of the sentences to be written into characters (that is, generates the spellings and produces the keystrokes or
character forms). The text capture process is the system (e.g. word processor, speech recognizer, or human transcriber) used to capture these characters and display them as text sentences.

In our studies we modeled the writing activity as a transcription activity in which the text to be entered is externally given. This simplification retains many of the aspects of the activity that are affected by K's dyslexia. The translating process in the Hayes and Flower model is replaced with a preordained text that the person must first read, encode, and store and then use to produce text for display on a computer screen or on paper. Transcription exercises memory, word recognition, lexical access, syntactic skill, spelling, and motor activation. It does not, however, exercise the generating, organizing, goal-setting, or translating processes of the composition task. Although performance in these is likely also to be handicapped by dyslexia, they are much more complicated to study and model and are left for future studies.

The models we construct are from the goms family introduced by Card, Moran, and Newell (1983). Goms models are derived from a Model Human Processor (MHP) that is defined by processors, memories, and principles of operation governing these. There are three processors: perceptual, cognitive, and motor which may work concurrently, subject to some data flow limitations. Each processor has a characteristic cycle time and other properties that are described by Card et al. Associated with the perceptual processor is a sensory memory for visual and auditory images. The cognitive processor operates in conjunction with working memory and uses information previously stored in long-term memory.

Goms models are based on a cognitive structure of the user consisting of (1) the user's Goals, (2) the basic Operations that the user can perform to achieve the goals, (3) Methods or combinations of the basic operations, and (4) Selection rules for choosing among alternative methods. The basic goms structure can be realized with different types of control structures, and operations can be defined at various levels of aggregation to match the grain of the data that are available and the purposes for which the model is being used. We represent behavior by a flow graph rather than by a program and use operators that are at a fairly high level of aggregation.

In our flow graph models, each observed or inferred process is represented by a rectangular box. Shown beneath each box are durations, rates, ranges of such measures, and other parameters of the process, such as chunks of working memory consumed. Directional lines connecting the processes indicate the observed or inferred flow of the activity. (This is not all the possible flow of the activity, which might be from any process to any other process.) The video recordings provided a detailed record of each subject's performance, from which we estimated the durations and other properties of most of the processes represented in the models. Where direct observation of process parameters was not possible, we inferred the durations from related experiments or from recourse to the literature.
The flow models enable us to estimate path times for both observed and unobserved routes through the model. From these data we are able to estimate worst case and best case rates for the writing methods and devices used. From the worst case analyses we can identify bottlenecks. This leads us to propose designs that relieve the bottlenecks. The models can then be used in the opposite direction. We restructure the model in accordance with the proposed design and reestimate the parameters of the new model. Redoing the best and worst case analyses then enables us to compare the observed activity with the way that that activity might develop in the new setting, utilizing the new design. We must be careful in carrying out this analysis to be sure that we do not introduce new bottlenecks in the designs.
Method of estimating parameters

Parameters of the models were obtained in one of three ways: by direct measurement from the videotapes, by borrowing from the literature, or by interpolation between observations and borrowed values. Because we are dealing with a special population, we obviously preferred direct measurement, but this was sometimes not possible. When we were forced to rely on other work, we carefully chose what to borrow and borrowed only as a fallback. When it was possible both to measure and borrow parameter values, we tried to compare them. Conflict between measured and borrowed parameters may point up interesting differences between dyslexics and nondyslexics, and we note these in the following discussion.¹

Our method of direct measurement involved three steps. First, we selected a part of the video in which there was continuous action of the sort that we deemed typical of this activity. Next, we carefully transcribed the observable action into a state diagram. From this diagram we could read off the times at which most of the overt actions (e.g. key strokes and head movements) take place and their durations. Finally, simple computations lead from these data to an average value for the desired parameter.

It is important to realize a significant limitation of this process and of the present effort more generally. There are many potentially important steps of the activity that are not distinguishable by our method. If there was good reason to assume more steps than we observed, we could use various mathematical means to estimate their durations, but we rarely encountered situations in which this seemed necessary in order to obtain a model that was satisfactory for our present purposes. Where this need is felt most strongly is in the case of such mental processes as spelling and in such visual activity as searching for one's place.

The reader probably shares some dissatisfaction with the authors in these cases, as the models seem to compress very complex activities into rectangular boxes. This diagrammatic representation elides a great deal of information, such as what the internal and external representations of information are, how they are translated one to another, where and how internal information is stored, and how branching decisions are made. Much of cognitive psychology is, of course, concerned with unpacking these boxes, and there is certainly literature that we could go to on this point. However, in accordance with our earlier hesitation to rely upon "mean-valued" science to provide useful data on exceptional cases, we choose to rely upon our own data to give outward constraint to these composed complexes and then work back from there at our own pace.

¹ In the model diagrams we indicate the source of a parameter. Values with no notation were directly measured. Values with a lowercase letter superscript were borrowed from another source. (The sources are displayed in the table legend.) Underlined values were interpolated.
A model of expert transcription typing in flowchart form adapted from John's (1988) program structure model. Upper flow model (a) has the parameters used by John to represent the performance of her subject. Lower flow model (b) has the parameters for nondyslexic subject S. S's typing rate computed from the model is 48 words/minute. The observed rate was 47 words/minute. Parameters with superscript j were borrowed from John (1988). Others were obtained from measurements made on video recordings.
Models for transcription typing

The first model, depicted in Figure 7.2a, is adapted from John's (1988) study of skilled touch typing. Chunks consisting of words, syllables, or letters are extracted from the source document (three chunks at a time in 0.34 second); the spelling of a chunk is obtained, unless it is a letter (0.05 second); there is a letter initiation operator (0.05 second); and then the letter is typed. For a slow typist (30 words/minute) the last step takes 0.23 second; for a fast typist (60 words/minute) it takes only 0.07 second. Loop transitions then take the person back either to get new chunks from the source document or to initiate the next letter. John's model allows parallel operation by the perceptual, cognitive, and motor processors, so, for example, a new chunk can be acquired while characters are being typed.

This model is a good representation of transcription typing by our nondyslexic subject, S (Figure 7.2b). To model her performance we borrowed the chunk acquisition, spelling, and letter initiation times from John's model and used the observed letter typing time of 0.15 second. We also observed from handwriting transcription experiments with S that she appeared to acquire four chunks (mostly words) at a time rather than the three used by John. When we employ these parameters in the model, we compute a typing rate of 48 words/minute (assuming text acquisition and typing to be concurrent), which compares with an observed rate excluding errors of 50 words/minute and a rate including errors from Table 7.1 of 47 words/minute.

We next turn to a slightly more complicated model of dyslexic transcription, "sight" (looking at the keyboard) typing (Figure 7.3). The main structural difference between this model and the model adapted from John (above) is the inclusion of a head movement operator (0.44 second to look over to the source document and back to the keyboard). As for differences in duration (and capacity), we see that K acquired one chunk with each look at the source and that acquisition took 1.1 seconds (time to reacquire position in text is included here), whereas John calculated three chunks in 0.34 second for nondyslexic typists. We inferred from John's model that the cognitive processor time to get the spelling of the next chunk was 0.05 second. Finally, the type-letter operator was observed to be 0.5 second for K versus 0.07 to 0.23 second for John's nondyslexic subjects and 0.15 second for S.

The word rate computed from the model was 10 words/minute (about 1 second/character), compared with the observed rate of 12 words/minute (Table 7.1). The combination of head movement time, the long times to acquire text and type the next letter, and the fact that only one chunk is acquired at a time all contribute to K's slow typing rate. The limited number of chunks that are acquired, a major impediment to increasing speed, reflects either working memory impairment of this subject or the fact that working memory is required for the type-letter operation (it is not automated).
The speech recognition experiments were carried out with the DragonDictate system operating on a 33 MHz 80486 PC-compatible computer. It is a trainable, word-at-a-time speech recognition device for simple transcription input to a text editor and other applications software. The system has a 30,000-word active vocabulary. The user wears a headset with one earphone and a microphone. In a training session the DragonDictate software asks the user to speak a training vocabulary that includes the special words used in editing (described below). After this the user can enter and edit arbitrary text using only the speech recognizer, however, he or she needs to pause briefly between each word. The system adapts to the user's speech patterns and to the user's vocabulary during use. Adaptation is accomplished by requiring the user to confirm or correct the
recognition decisions made by the system and to enter new words into the vocabulary by typing them after they have been spoken. This allows the system to associate a speech pattern with the new word.

The process of text entry goes as follows. The user speaks a word which the DragonDictate attempts to recognize. It appends its best guess for that word to the line of text on the computer screen. If it is the intended word, the user can simply proceed and speak the next word. Other guesses for the word are shown on the screen in an ordered short list called a choice menu. If the word added to the text line is not correct and the correct word is on the choice menu (say, the number 4), the user can say, “choose four,” and that word will be entered on the text line in place of the erroneous word. Alternatively, the user can say, “Scratch that” (one of the learned command phrases) to cause the erroneous word to be removed and can repeat the word in the hope that it will be recognized correctly the second time. If the correct word is not on the choice menu, the user can start typing the word. If the word is in the system’s vocabulary, the first few letters will usually cause the appropriate word to appear on the choice menu. If the word is not in the vocabulary, the user must type it completely and press the enter key or give the command “OK” to cause it to be appended. (There is also an editing mode, which we will not try to describe here, as it is not relevant to the present discussion, is complex, and was hard for K to use.)

**Parameter estimation**

Figure 7.4 is a partial model of K’s use of the DragonDictate. The reason this model is only partial is that there are pathways that K took only once, or a very few times, or which were so long and tortuous that we have simply removed them from consideration in our analyses (e.g. editing). Also, this model represents K’s final level of performance in our data, skipping over the training phase (of which the very different structure presumably has little bearing on use). More important, this model does not represent the use of the device that one might come to as an expert – but, as we shall see, this will also have little bearing on our analyses.

The uppermost part of the model is the speaking loop, analogous to the typing loop in the other models. The initiate-and-speak-syllable operation required 0.2 second per syllable. This version of the DragonDictate software required about 0.42 second to recognize and append a word to the text line and 0.22 second to put up the choice menu. Note that this model includes a read-and-check-recognized-word operation (0.27 second), which was unnecessary in the other models. Usually K acquired only one chunk (a single word) when looking at the source text. Sometimes, if the text consisted of short, easy-to-pronounce words, he might acquire two chunks (also words). We observed that after confirming that the first word was correct (read-and-check operation) there was a latency of 0.5 second before the second word was spoken. We have identified this latency
words into the
book which the
user can
read
for that word to
be correctly
recognized.

The reason this
was removed
was because
of a very
unusual
case in
which the
user pressed
the space
bar instead of
typing the
word. This is
not relevant
to the
model
in
Figure 7.4. Partial flow model of K's
use of the DragonDictate speech
recognizer for transcription. Parameters with superscript c were
obtained from an experiment in essay
composition performed with K at the
same time as this transcription experiment. Other
parameters were obtained from the
transcription video recordings.

Loop time for 2 syllables/word if
correct recognition of word = 3.0
sec/word; 20 words/min
Loop time if “choose n” required = 6 sec/word; 10 words/min
Loop time if typing 2 letters required and then “choose n” = 11 sec;
6 words/min
Overall rate from loop times = 13 words/min
Actual transcription rate = 9 words/min

with the get-pronunciation operator shown in Figure 7.4. We attribute this rather
long latency to the care K was taking in trying to pronounce words correctly and
consistently and in making sure that there was a gap between words.
The lower part of the figure represents the "choose n" and type-letter error recovery paths. We observed the menu search time to be about 2 seconds. The other operators involved in these paths and their times are shown in the figure. It should be pointed out that there was great variability in the time to type words or even the first few letters of a word. If the word was easy, so that its spelling was well known and easily retrievable by K, it would take about 0.6 second to get the spelling and 0.4 second to type each character. Since K is not a touch typist, he had to make a head movement from the screen to the keyboard in order to type. If the word was difficult and he had to obtain the spelling from phonological and lexical rules or from repeated reference to the source, successful transcription of the word could take a long time and involve several traverses of the typing-read-and-recognize loop. We observed an occasion when he spent almost 30 seconds on one word.

**Loop times**

These models enable us to compute loop times—the average time it takes the person to cycle through various pathways in the model. From loop times one can easily compute rates in terms of words/minute that are meaningful to the individual user and useful in comparisons. If the model is of high fidelity, then we can use such times to find best and worst case paths and thus to compare expected performances under different conditions.

Three loops are shown in Figure 7.4. The main loop at the top of the figure is for words that are recognized correctly; it takes about 3 seconds/word (assuming a word of two syllables) to read the word, speak it, wait for it to be recognized and posted on the display, and to read and check the correctness of the displayed word. This corresponds to a rate of 20 words/minute. The second loop, shown in the middle of the figure, is for cases of a word that is incorrectly recognized but appears on the choice menu. This path takes about 6 seconds/word (10 words/minute); the additional time goes to finding the word in the choice menu, speaking "choose n," waiting for the word to appear in the text line, and checking it. The third loop, at the bottom of the figure, is for cases where the word is incorrectly recognized, does not appear in the choice menu, and must be typed. For the case in which two letters are typed and then a "choose n" loop is traversed, the loop time is about 11 seconds/word (6 words/minute). We observed that the main loop was taken for about 67% of the words and each other loop for about 16% of words. This gives a composite performance of 13 words/minute compared with the observed performance (Table 7.1) of 9 words/minute. The difference between model and actual performance can be attributed to the time taken to correct dictation (as opposed to recognition) errors or to deal with complex words that are not included in the model.
Model of collaborative transcription with human transcriber

Table 7.1 showed that when K was working in collaboration with a human transcriber he was able to achieve rates of 69 words/minute (after subtracting time spent waiting for the typist). Figure 7.5 shows the flow model for this method of transcription. The initiate-and-speak-syllable operation was observed to take 0.2 second. We used John's value for the cognitive processor time of 0.05 second for the get-pronunciation time. We observed from K's speaking patterns that he acquired text in one-word, two-word, or three-word chunks with about equal probability. We could not observe the chunk acquisition time directly and used a value of 0.7 second that is an interpolation between John's value for acquisition time (0.34 second) and the acquisition time (1.1 seconds) observed in the typing transcription experiments (Figure 7.3) and that gives good agreement with the observed rate. The word rate predicted by the model is 68 words/minute versus an observed rate of 69 words/minute.

Possible design improvements in speech recognition prostheses

The method of comparative modeling enables us to answer the question of why the speech recognition prosthesis is not as useful as expected for K's online writing. Furthermore, it enables us to suggest design changes that will improve performance for users with characteristics similar to those of our dyslexic subject. Performance with the speech recognition prosthesis was observed to be 9 words/minute versus 69 words/minute when working collaboratively with the hypothetical infinitely fast typist. The comparison between the two models is 13 words/minute for speech recognition versus 68 words/minute for collaborative transcription. Even the rate for correctly recognized words (the fast loop in the operation of the speech recognition device, is only 20 words/minute.) The two incorrect-recognition loops lead to significant decrements in the performance, down to 13 words/minute. Finally, errors made by the user and difficult words (not in the model) further reduce performance.

Clearly the collaborative setting is an easier and more natural environment for K. The cognitive demands associated with the actual production of text, with device control, and with error management are less than with DragonDictate. Can we look to the collaborative setting for guidance on how the DragonDictate system might be modified to give faster word production rates?

Adjust recognition speed

There are several major differences between the collaborative and the DragonDictate settings. The data presented for collaborative transcription assumed that typing takes zero time, whereas it takes 0.42 second for DragonDictate to recognize a word and enter it on the text line. To put the
Figure 7.5. Flow model of collaborative transcription in which K dictated text to a human transcriber. Time K spent waiting for typist to finish previously dictated material has been subtracted from total time. This is equivalent to assuming that typing a character takes 0 seconds. Parameter with superscript \( j \) was borrowed from John (1988). The underlined parameter (time to acquire next chunk) was interpolated as described in text.

Loop time for 1-word chunks = 1.15 sec/word; 52 words/min
Loop time for 2-word chunks = 0.8 sec/word; 75 words/min
Loop time for 3-word chunks = 0.68 sec/word; 88 words/min
Overall rate of text production from model = 68 words/min; observed rate = 69 words/min

Comparison on an equivalent basis, we should examine the case in which recognition is instantaneous, so that the recognized word appears on the text line with no delay. If we make this assumption and set the word-recognition-and-display operator to zero, the word rate for the correct-recognition loop increases to 24 words/minute. When the error loops are included, the rate becomes 15 words/minute, still substantially below the collaborative rate of 68 words/minute.

Auditory notification of uncertain words

DragonDictate requires that each word be verified visually. This takes time for verification and requires head movement after every word to examine the computer display for correctness. There is no need for visual checking or for
head movement in the collaborative setting, because the typist simply types on
when there is no uncertainty and gives signals to indicate confusion, mishearing,
or spelling difficulties, thereby leaving the reader's attention for the most part
uninterrupted. The recognizer might give a similar kind of feedback. For
example, an acknowledgment sound (perhaps the sound of keystrokes) might be
used to indicate recognition that has a high probability of being correct. When
there is less confidence that the word is correct, the recognizer might read the
word back to the user with a questioning intonation and wait for an "OK"
command. When there is even more uncertainty, the recognizer might signal
with a "Huh?" If this kind of feedback were given by the DragonDictate, the user
would not have to check words that have a high probability of being correct, and
the head movement time of 0.52 second/word and the read-and-check operation
(0.27 second) would not be necessary. In addition, the get-pronunciation latency
would probably be reduced, because speech rhythm would be less often
interrupted by the need to attend to other tasks. We will assume that it is reduced
to 0.2 second, about halfway to the value of 0.05 second used by John. Finally,
the acquisition time would probably be reduced, because the reader would not
lose his or her place in the source text and have to relegate it. We might assume
that the acquisition time under these circumstances would be the same as when
dictating to a human typist, where a value of 0.7 second gave a good fit to the
data (see Figure 7.5).

Revised flow diagram

We have redrawn the flow model for the DragonDictate to reflect these changes
(Figure 7.6). Note that the correct-word loop is much simpler and shorter. These
changes would increase the correct-word rate to 43 words/minute, a value that is
reasonably close to the 68 words/minute obtained from the model for
collaborative transcription. However, the aural feedback generates the need for a
correction loop to confirm an uncertain recognition with an "OK." We assume
that confirmation will take 0.4 second, the time observed for speaking two
syllables. When the user receives the signal indicating an incorrect recognition,
he or she must make a head movement (0.26 second) to the display to see if the
word is in the choice menu. We assume that incorrect recognitions occur with the
same probability as before (16% requiring "choose n" and 16% requiring typing)
and that half of the correct recognitions (34%) are certain, requiring no action on
the part of the user, and half (34%) require a confirming "OK." We do not have a
better basis for estimating the probability of traversing these loops. Given these
assumptions, the overall rate computed from the model is 19 words/minute.
Figure 7.6. Postulated flow model of K's use of the DragonDictate speech recognizer that has instantaneous recognition and is redesigned to use aural feedback to indicate the level of uncertainty in the recognition of words. Parameters are mostly drawn from Figure 7.4. Underlined parameters are interpolated as described in the text.

Loop time for correct recognition = 1.3 sec/word; 43 words/min
Loop time for uncertain correct recognition = 1.7 sec; 35 words/min
Loop time if “choose n” required (no check of recognition of “choose n”) = 4.7 sec/word; 13 words/min
Loop time if typing 2 letters required and then choose n = 8.9 sec; 7 words/min
Overall rate from loop times = 19 words/min assuming correct signal 34%; uncertain word 34%; choose n 16%; typing 16%
Further increases in word rate would require acquisition of multiple chunks (words), which is certainly possible, but we have no basis for determining whether it will happen, given the requirement for pausing between words and for paying some attention to signals indicating errors. In any case, we have moved well beyond our actual experimental data, and further extrapolation seems unwarranted.

Error rates

It should be noted that although we have drawn designs to improve the primary correct-word loop, we have done nothing yet to improve the error loops. Left unchanged, they limit the overall performance to 19 words/minute. To increase the overall rate substantially, it is necessary either to decrease the fraction of words that are incorrectly recognized or to decrease the time to handle these errors. Error rates should decline with additional training of the system to the user's voice patterns and vocabulary. A reduction of error rate by half from 32% to 16% would cause the overall transcription rate to increase to 26 words/minute. We do not have data for estimating how long it would take to achieve these levels, but representations by Dragon Systems suggest that they are achievable by nondyslexic speakers.

Correcting errors

As we have seen, error handling has a major effect on overall performance. We have two suggestions for improvements in this area. The first is to process errors on a sentence basis so that the user can dictate an entire sentence without shifting attention back and forth from the writing task to the correction task. The measures derived from the spoken words of the entire sentence would need to be buffered. Upon completion of the sentence, the choice menu for the uncertain words would be displayed, and the user could proceed to make the required "choose n" or "type letter" corrections in a second pass through the sentence. The primary advantage of this method is that it will allow the user to dictate a complete thought before being interrupted and thereby speed up the dictation process. It might also result in less time being spent on each error, but further experimentation would be required to confirm this.

A second type of error situation, not represented in the model but mentioned in the discussion, is the problem in correcting errors in words that the user finds difficult to spell. We observed situations in the essay composition experiment in which DragonDictate could not recognize the spoken word and the subject could not spell it. K sometimes spent an extraordinary amount of time, often several minutes, trying to find some way of getting the correct word either by persevering in his attempts to spell it or by speaking similar-sounding words that the system might recognize correctly and that he could modify to the desired
word. At times this consumed 20% to 40% of the total time required to capture the text. An escape from this kind of deadlock is needed. A recorder mode in which the speech for these words is recorded for later resolution, or a syllable mode in which the user speaks the phonetic components of the word that are then recognized and resolved later, are possible approaches. There is a significant opportunity for performance improvement if this situation is handled more effectively.

Conclusions

Dyslexia is a relatively common condition that can produce to significant disability. Dyslexics can be especially handicapped by the influence of human factors models that aim to describe a generally nondyslexic population. We sought to remedy this situation by developing detailed models of a particular dyslexic person in computer supported writing and by using these models in the analysis and design of systems.

In studying dyslexic and nondyslexic individuals in a variety of computer aided transcription settings, we found that a speech recognition prosthesis did not compare well in terms of word production rates with collaborative writing, nor was it very easy to use. It did, however, compare reasonably well with typing and handwriting, and the fact that it eliminated most of the burden of spelling was seen as a significant benefit.

Guided by a general model of the writing process, we produced detailed models to describe these situations. The models helped us to understand the bottlenecks in computer aided writing that lead from dyslexia to significant handicaps. Three types of bottlenecks account for most of the performance degradation:

1. operations in which the subject had to confront his areas of impairment, such as spelling words, required excessive times to complete;
2. dealing with errors in the operation of the speech recognition computer system used to support writing, such as correcting errors in recognition of spoken text or command words, accounted for a large fraction of the total time spent doing the primary writing task;
3. the cognitive load of controlling the speech recognition system, whose considerable complexity is depicted by the multi-path flow diagrams shown earlier in the chapter, interfered with performance of the primary writing task.

These three types of bottlenecks are likely to be very important to the performance of all computer-based systems designed to assist people with disabilities. It is important that computer aids be designed to minimize their effects.
The detailed models developed here illustrate a general method for determining where and how computer-based aids should be redesigned so that good performance is achieved and performance bottlenecks are avoided. Analysis of the model of speech recognition transcription enabled us to propose several ways in which the speech recognition system might be modified to give better support to people with dyslexia. The collaborative setting, where the differences in performance are not so great, provided insights about what kinds of design changes would be most effective. These changes include various modes of audible feedback to reduce the time spent in visually checking words that are correctly recognized, changes in the general dynamic of interaction so that errors can be dealt with on a sentence-by-sentence instead of word-by-word basis, and a novel means of correction that uses syllabic input for difficult words. These changes should reduce cognitive load on the writer and thus enable him or her to allocate more attention to the primary task of composition, a task that dyslexics find difficult in any form. Finally, we were able to use the models to estimate the improvement in speech production rates and error frequency that should result from the proposed changes.

The method of detailed observation and modeling and especially of the analysis of natural successful settings is general enough to be used with individuals with widely different capabilities and in a large range of situations. We have shown, moreover, that it can provide a rational basis for individualizing system design. Broad application of this method to different persons with different capabilities and in a variety of settings will make available a large collection of relatively specific models. A library of such models may aid both system designers and therapists concerned with prescribing devices to assist dyslexics in bridging the gap between individuals and their goals.

Acknowledgments

The authors would like to thank Stuart Card for his contributions to planning the experimental studies and developing the models presented in this paper and Becky Agin and Jeannine Pinto for their helpful comments about earlier drafts of this chapter.

References


