Accuracy Error Span Evaluations: Implications for User Interface Design of Speech Dictation and Handwriting Recognition Applications

James R. Lewis

IBM Pervasive Computing Division
8051 Congress Ave, Suite 2227
Boca Raton, FL 33487
jimlewis@us.ibm.com

Abstract

This paper describes a number of experiments and analyses conducted to gain an understanding of the typical properties of error spans – the number of consecutive incorrect words generated when producing text with recognition technologies. The data were drawn from three studies of recognition applications: one study of an IBM dictation system, one of a non-IBM dictation system, and one of an IBM handwriting recognition application. Across the three recognition studies, the appropriate width for a text field designed to display text selected for correction appeared to be the width that would accommodate the presentation of three words (or 17 characters). An analysis of the position of correct text in the alternates list for the IBM dictation system indicated that a correction control should ideally present four alternates (covering 97% of the cases for error spans of one word and 90% of cases for error spans of two words). There appears to be little reason to provide a control that displays more than four alternates for any type of recognition user interface.

1 Introduction

Research in the component tasks of producing text via recognition technologies based on speech and handwriting input has shown that in many cases the speed of making corrections can have a greater influence than recognition accuracy or input rate on the throughput of text production (Commarford & Lewis, 2004; Karat, Halverson, Horn, & Karat, 1999; Lewis, 1999). The analyses in this paper explore the properties of error spans – the number of consecutive incorrect words generated when producing text with recognition technologies. These analyses can inform the design of correction procedures, with the potential effect of decreasing the time required to correct recognition errors.

Table 1 shows examples of error spans from one to four words (with the error span appearing in bold text). Note that the error span counts are for the output of the recognizer – not the input. For speech systems, it is common for the input to only be available as an audio file for a limited period of time. The input for handwriting systems can either be temporary (if written on a device that does not preserve the handwriting, such as the handwriting input area on a PDA) or permanent (if written using a product such as a CrossPad[1] with IBM[2] Ink Manager[3], which was designed primarily for handwriting capture and only secondarily for handwriting recognition).

[1] CrossPad is a trademark or registered trademark of Cross Pen Corp

Table 1: Examples of Error Spans

Span Length	Input	Output
1	They played a game of soccer yesterday.	They played a game of **saga** yesterday.
2	He lived in Mochrie until 1986.	He lived in **Murray Creek** until 1986.
3	She would frolic in the park.	She would **for all like** in the park.
4	Our new product is Voice Express Plus.	Our new product is **voice six glass loss**.

It is possible to tune recognition correction procedures to optimize them for different error span lengths. Thus, it is important to understand the relative frequency of occurrence of error spans as a function of length. The property can also influence the appropriate size of text controls for displaying a list of correction alternates after a user has selected a span of errors for correction. This is particularly important when designing user interfaces for devices with small displays, such as PDAs.

The following analyses use data from three studies of recognition accuracy. Two were dictation studies (one with an IBM product, one with a non-IBM product) and one was a handwriting recognition study (using an IBM handwriting recognizer).

2 Experiment 1: IBM Dictation

2.1 Method

The purpose of this experiment was to measure (1) the likelihood that misrecognition errors span more than one word and (2) for each observed error span, the likelihood that the desired alternate is in the correction dialog list.

Four participants (employees of IBM) provided samples of dictation, speaking to the IBM ViaVoice[4] '98 Refresh system running on a computer with the Microsoft Windows[5] 95 operating system. After dictating a paragraph, participants proofread the dictated text and recorded in a spreadsheet:

- The number of words in the source document for each misrecognition error
- The error span (the number of words produced by the recognizer for that error)
- The number of alternates provided in the correction dialog for that error span
- The position of the correct alternative (if present) in the correction dialog
- The misrecognized word or phrase (as it appeared in the source document)
- The word or phrase produced by the recognizer for that misrecognition

In addition to enabling the measure of the relative frequency of occurrence of different spans of error, the data also permit the computation of a new measure of correction effectiveness – the

---

[2] IBM is a registered trademark of International Business Machines Corp.
[3] Ink Manager is a trademark or registered trademark of International Business Machines Corp.
[4] ViaVoice is a registered trademark of International Business Machines Corp.
[5] Microsoft and Windows are trademarks or registered trademarks of Microsoft Corp.

overall percentage of in-list correctable words. The logic of this measure is that when a user makes a multiple-word correction, they are correcting more words at once than if they attempted to correct each word individually, resulting in reduced user interface overhead. For example, suppose a user sees a three-word error. Correcting each word individually would require three separate selections, three separate examinations of the alternates list, and three separate decisions to accept an alternate or to type/redictate the correction, as opposed to the more efficient strategy of selecting all three words and making a single examination of the alternates list for the correct alternate. To calculate the overall percentage of in-list correctable words (presented in the alternates list), multiply the number of cases for a specific error span by the length of that span. If there were ten three-span errors, and four of those had the correct alternate, then the total number of words associated with this error span would be 30 (3x10) and the number of in-list correctable words would be 12 (3x4). The total number of in-list correctable words times 100 divided by the total number of correctable words produces the new measure (also called the in-list correctable word rate). These data are based on the production of over 5000 words across all the participants.

2.2 Results

Table 2 shows the error span measurements for this experiment. The In List column contains the percentage of correctable words that were in the alternates list after selecting the indicated error span. The Out List column contains the percentage of correctable words that were not in the alternates list. Note that the entries for In List and Out List at each level of Error Span sum to 100%. The Total column indicates the percentage of error spans that had the indicated length, and the Cumulative Percentage column shows how the total percentages as a function of error span led up to the grand total of 100%. Inspection of the Cumulative Percentage column shows that 78% of error spans had a length of 1, and 99.1% of error spans had a length of 3 or fewer errors. The longest observed error span had a length of five words.

Table 2: Span Measurements for IBM Dictation Experiment

Error Span	In List	Out List	Total	Cumulative Percentage
1	51.7%	48.3%	78.0%	78.0%
2	26.6%	73.4%	17.0%	95.0%
3	15.8%	84.2%	4.1%	99.1%
4	0.0%	100.0%	0.4%	99.5%
5	0.0%	100.0%	0.5%	100.0%

3 Experiment 2: Non-IBM Dictation

3.1 Method

The method for this experiment was identical to that of Experiment 1 with the exception of the product used to recognize the speech input. In this experiment, the product was one of IBM's competitors[6].

[6] The name of the competitive product is not relevant for the purposes of this paper. The data are included to enhance the generalizability of the final conclusions.

3.2 Results

Table 3 shows the error span results for Experiment 2. The In List column contains the percentage of correctable words that were in the alternates list after selecting the indicated error span. The Out List column contains the percentage of correctable words that were not in the alternates list. Note that the entries for In List and Out List at each level of Error Span sum to 100%. The Total column indicates the percentage of error spans that had the indicated length, and the Cumulative Percentage column shows how the total percentages as a function of error span led up to the grand total of 100%. Inspection of the Cumulative Percentage column shows that 73.3% of error spans had a length of 1, and 98.3% of error spans had a length of 3 or fewer errors. The longest observed error span had a length of six words.

Table 3: Error Span Measurements for Non-IBM Dictation Experiment

Error Span	In List	Out List	Total	Cumulative Percentage
1	37.0%	63.0%	73.3%	73.3%
2	34.7%	65.3%	17.6%	91.0%
3	13.3%	86.7%	7.3%	98.3%
4	25.0%	75.0%	1.0%	99.3%
5	0.0%	100.0%	0.5%	99.8%
6	0.0%	100.0%	0.2%	100.0%

4 Experiment 3: IBM Handwriting Recognition

4.1 Method

Four IBM employees participated in this study. Two of the participants were male and two were female. To collect the handwriting productions, a CrossPad with special software acted like a handwriting capture tablet in absolute mode. While in this mode, participants wrote the production scripts one at a time on lined paper with the recognition area marked out. This handwriting production was captured and passed through Ink Manager for recognition analysis.

4.2 Results

Table 4 shows the error span results (based on the production of 320 words across all participants) for Experiment 3 (Totals and Cumulative Percentages only – at the time this experiment was conducted, Ink Manager did not have a correction dialog, preventing the measurement of correctable words in the alternates list). The Total column indicates the percentage of error spans that had the indicated length, and the Cumulative Percentage column shows how the total percentages as a function of error span led up to the grand total of 100%. Inspection of the Cumulative Percentage column shows that 83.1% of error spans had a length of 1, and 99.7% of error spans had a length of 3 or fewer errors. The longest observed error span had a length of four words.

Table 4: Error Span Measurements for IBM Handwriting Recognition Experiment

Error Span	Total	Cumulative Percentage
1	83.1%	83.1%
2	14.7%	97.8%
3	1.9%	99.7%
4	0.3%	100.0%

5 Additional Analyses

5.1 Error Span Percentages across the Three Recognition Experiments

Figure 1 shows a graph of the cumulative percentage of error spans for the three recognition experiments.

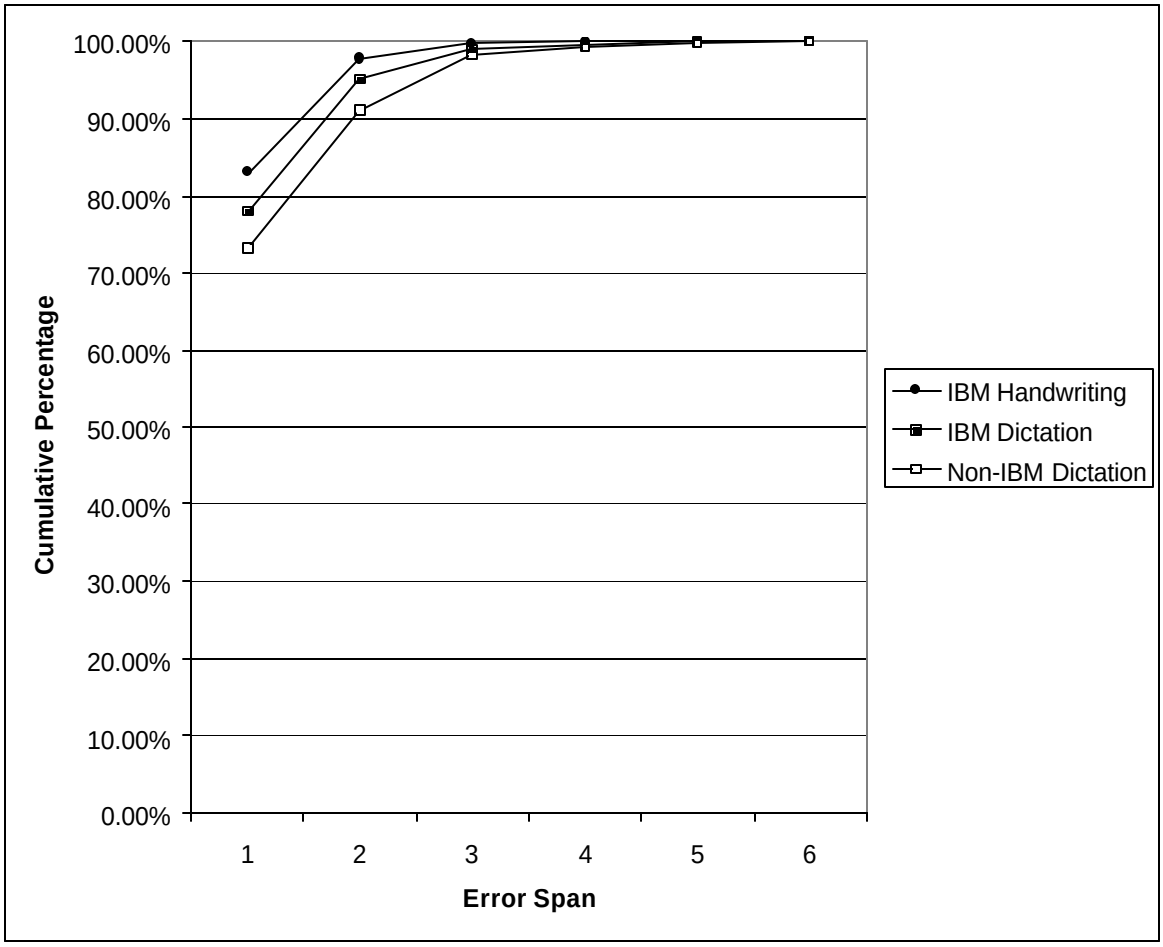

Figure 1: Error span percentages across three recognition experiments

Although not identical, the error span curves are very similar. In all three cases, the majority of error spans were one word. Very few error spans had a length greater than three words.

5.2 Analysis of Position of Correct Alternate in Alternates List

This analysis used data from the IBM dictation experiment to estimate how many correction alternates a dictation application should provide in a product with a small display. Table 5 shows, for each error span, the average number of alternates provided in the correction dialog, whether any of the alternates in the alternates list provided the correct text, and, if the correct text was provided, its position in the alternates list. The columns for position in the alternates list stop after Pos 5 because the correct alternate was never present after the fifth position.

Table 5: Position of Correct Text in Alternates List

Error Span	Number of Alternates	Not Listed	Listed	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5
1	4.6	124	151	93	27	17	10	4
2	3.3	64	10	0	6	2	1	1
3	0.6	16	0	0	0	0	0	0
4	0	2	0	0	0	0	0	0
5	0	4	0	0	0	0	0	0

For error spans of three or more words, the correct text never appeared in the alternates list. For an error span of two words, the correct text appeared as an alternate on ten occasions, with six of those in the second position, two in the third position, and one each in the fourth and fifth positions. From Table 5, given the denominator of 10, the cumulative percentage at Position 3 for an error span of two words was 80%, the cumulative percentage at Position 4 was 90%, and only one case occurred at a position greater than 4.

There were 151 cases in which the correct text appeared as an alternate for error spans of one word. Table 6 shows the percentage and cumulative percentages for this error span as a function of position in the alternates list.

Table 6: Percentage and Cumulative Percentage of Presence of Correct Text as a Function of Position in the Alternates List for Error Spans of One Word

Position	Percentage	Cumulative Percentage
1	61.6	61.6
2	17.9	79.5
3	11.3	90.7
4	6.6	97.4
5	2.6	100.0

Figure 2 provides a graph of the cumulative percentage for the presence of the correct text (if provided at all) as a function of position in the alternates list for error spans of one word.

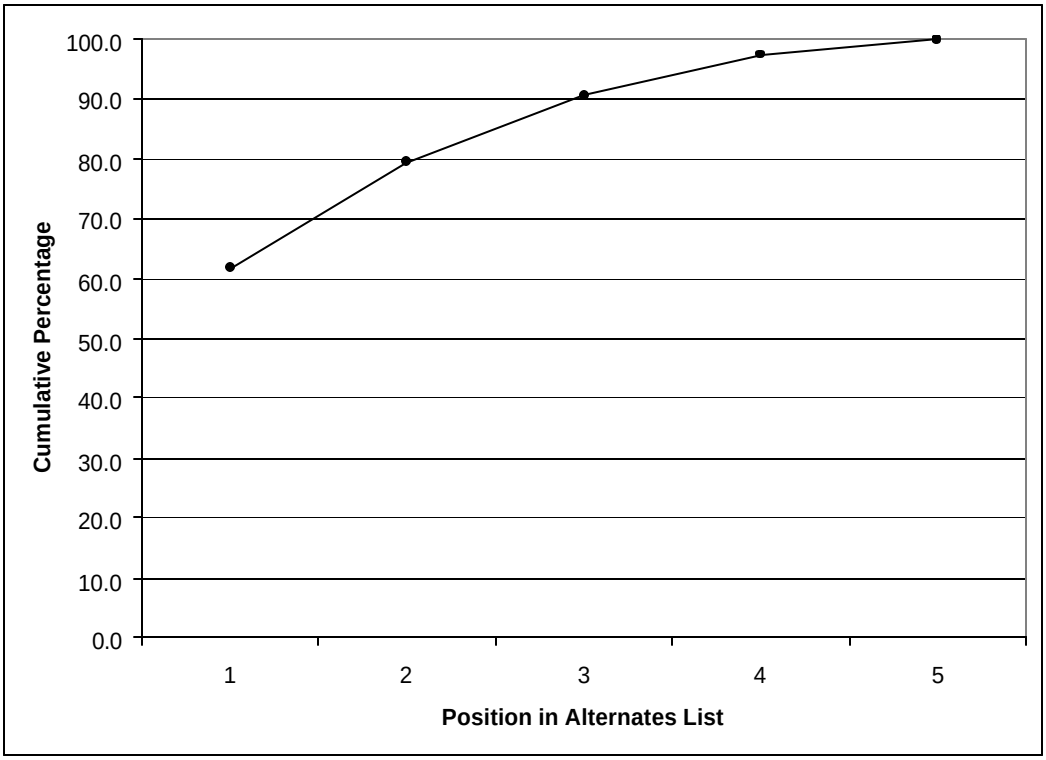

Figure 2: Position of correct text as a function of position in the alternates list for error spans of one word

6 Discussion and Design Recommendations

The reported analyses have implications for two aspects of the design of dictation and handwriting recognition applications: (1) the width (horizontal) of the text field used to display text selected for correction and (2) the height (vertical) of the control used to display the correction alternates.

6.1 Recommended Minimum Width of Field for Displaying Text Selected for Correction

Across the three recognition studies, the appropriate width appeared to be one that would accommodate the presentation of three words. Given an average English word length of five characters, this suggests that the width of the field under discussion should be no less than 17 characters (three words plus two spaces). If display space permits, then the field should be correspondingly wider. If the space does not permit a field of this width, then there will be times when users must scroll horizontally to view the entire alternate.

6.2 Recommended Minimum Height of Control for Displaying Correction Alternates List

The analysis of the position of correct text in the alternates list indicated that this control should ideally present four alternates (covering 97% of the cases for error spans of one word and 90% of cases for error spans of two words). There appears to be little reason to provide a control that displays more than four alternates for any type of user interface. If the display space does not permit the display of four alternates, then the data presented show how this will affect the ease of selecting an alternate. For example, if the control shows three alternates, then this will cover about 90% of cases for error spans of one word and 80% of cases for error spans of two words. A control that showed only one alternate would cover about 60% of cases for error spans of one word and none of the cases for error spans of two words.

7 References

Commarford, P. M., and Lewis, J. R. (2004). Models of throughput rates for dictation and voice spelling for handheld devices. *International Journal of Speech Technology*, *7*, 69-79.

Karat, C.M., Halverson, C., Horn, D. and Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. *CHI 99 Conference Proceedings* (pp. 568-575). Pittsburgh, PA: ACM.

Lewis, J. R. (1999). Effect of error correction strategy on speech dictation throughput. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 457-461). Santa Monica, CA: Human Factors and Ergonomics Society.