Usability Study of Text Entry Interfaces for Personal Organizers

Isabelle Guyon[1], Yann Le Cun and Colin Warwick
AT&T Bell Laboratories
Holmdel, New Jersey, USA

Abstract:

We performed comparative tests of several keyboards and handwriting recognizers to determine their usability as text entry interfaces. For very small devices (e.g. cellular phones), the keyboard is impractical. alternative user interfaces need to be invented, but handwriting recognition is marginally usable For medium size devices (e g palmtop personal organizers), small "hardware" or "software" keyboards are both most efficient and preferred by users Ease of correction was found to be at least as important as recognition accuracy Users are biased in favor of handwriting, even if it is less efficient, particularly when the size of the keyboard is so small that typing is frustrating The handwriting recognition system designed in our department beats other recognizers

Keywords: Pen computers, Personal organizers, Cellular phones; Wireless communicators, Personal communicators, PDA, Palmtop computers, Handwriting recognition

1 Motivations and Problem Description

Yesterday's phones were simple communication devices which could let you dial a number and would transmit your voice. The current trend in telecommunications is to integrate in the same device many communication modalities, including voice, fax and email. It becomes difficult to draw the line between phones and computers, as phones start incorporating agendas, calendars, calculators, etc. Examples of such devices include the Apple Newton personal communicator and the IBM/Bell-South Simon cellular phone. For a review see [4].

The success of personal communicators depends upon several factor including good hardware, good software, good telecommunication infrastructure and good user interface. The user interface factor which has been long neglected may end up being the major obstacle to success. Most personal communicators were designed around a handwriting user interface. This seemed a reasonable marketing strategy a few years ago, when everybody was ignorant about the performance of handwriting recognition. In light of the results obtained by handwriting recognizers, it is important to ask ourselves again today what are the chances that handwriting become a viable alternative to the keyboard.

Results reported in the literature indicate that users are demanding very high recognition accuracy. In previous studies [5, 2], we estimated that a character error rate of 1% is acceptable while 2% is untolerable. In this paper, we report the results of a comparative usability study test which factors in both the accuracy of the character entry device (handwriting recognizer or keyboard) and the ease of correction. Subjects are timed to enter the same sentence *without error* on several devices. Because it is such a boring task, the subjects are motivated to enter the sentence as fast as they can. However, if they go too fast, more recognition errors or typing errors will be introduced and the overall time might be worse because of the corrections. This task is more realistic than measuring error rates without allowing corrections because in this last case it is then unclear whether the subjects are motivated or not to write or type neatly.

[1]Current address: CyberGold, 955 Creston Road, Berkeley CA 94708 Email isabelle@cybergold net Phone 1 - 510 - 524 62 11 Fax 1 - 510 - 524 07 47

2 Experimental Setup

Twenty two subjects from AT&T Bell Labs research participated to the following test. We installed in our laboratory several user interfaces designed to enter (easily) text into personal organizers, personal communicators or cellular phones:
1 - CIC recognizer for handprinted characters.
2 - Newton (ParaGraph) recognizer of cursive/mixed handwriting.
3 - Graffiti recognizer (special alphabet of 1-stroke characters).
4 - Predictive keyboard of the Simon cellular phone (6 keys only displayed at a time).
5 - Small size "hardware" QWERTY keyboard of a Sharp Wizard.
6 - Very small size "software" QWERTY keyboard that fits on a cellular phone.
7 - LeRec handwriting recognizer (Bell Labs design [1]).
9- Sony Magic Link "software" QWERTY keyboard.

A more detailed description of the interfaces can be found in Section 3.

We also tested, for comparison, two reference text entry methods:
I - Regular QWERTY keyboard.
II - Pen and paper.

The subjects where all computer literates. All of them belonged to the technical staff of Bell Laboratories and had Master of PhD degrees, except for one secretary. Their age ranged between 30 and 55, with an average of 36. Their typing skills were variable.

The subjects could sit at a table. They were asked to get familiar with all the user interfaces by trying them out for a few minutes. Then, they could proceed with the rest of the test:

Step A: The subjects had to give a grade 0-10 corresponding to how usable they thought each interface should be to enter text (Grade before). Grades had to be given on all interfaces before proceeding to the rest of the test.

Step B: The subjects timed themselves writing the sentence (from Bob Lucky's book "Silicon dreams" [3]):

```
Common folklore is that the Dvorak keyboard is much better
than QWERTY and that if we could all retrain ourselves
we would come a long way ahead
```

The sentence has 27 words and 145 characters. The subjects were instructed to edit the sentence until they reached 0 error. They could choose to test the interfaces in any order, depending on availability.

Step C: The subjects were asked to grade 0-10 all the systems again (Grade after). The subjects filled out themselves a form with their results.

The purpose of asking the subjects to give to the interfaces a "Grade before" and a "Grade after" was to see whether performing the test would significantly change their perception of how usable the interfaces were.

This test measures "walk-up" user interface performance. After longer periods of training, users adjust to the interfaces and perform better. A "walk-up" performance test evaluates the reactions of a customer who enters a store and spends a few minutes testing various products before deciding to purchase or not. It is difficult (and perhaps not critical) to measure the performance of expert users because one cannot convince volunteers to spend much time training on bizarre interfaces. Since it is so boring and time consuming, paying subjects is expensive.

3 Description of the Devices

We briefly describe the devices used for the test.

Interface number	Mean grade before (Step A)	Mean time (sec.) (Step B)	Mean grade after (Step C)	Device
I	9 15	42.35	8 89	Regular keyboard
II	8 30	59 20	7 58	Pen and paper
5	6 35	98.24	6 57	Sharp Wizard keyboard
8	6 25	116 06	6.38	Sony Magic Link
7	4 64	256 36	4 80	LeRec (Bell Labs design [1])
3	4 05	396 25	3.84	Graffiti recognizer
6	4 84	208 47	3 79	Mini "software" keyboard
2	4 95	339 42	3 68	Newton cursive recognizer
4	3 86	294 40	3.61	Simon predictive keyboard
1	4 20	460 77	2 32	CIC handwriting recognizer

Table 1: Average values of the interface *subjective* grades and the average times for entering a sentence 27 word long, in order of decreasing "grade after"

3.1 Reference Interfaces

Interfaces I and II are used as points of reference. Interface I is a regular keyboard attached to a PC. Backspace is used to make corrections. Interface II is a pen and a paper. Subjects were instructed to write legibly.

3.2 Interfaces for Small Portable Devices

1. CIC handprint recognizer. A recognizer commercialized by Communication Intelligence Corporation (CIC). It handles handprinted characters that can be touching, but must be separated by pen-lifts. It uses a language model. "QWERTY" is a word of its dictionary.

   Subjects write on a 7 by 9 inch opaque tablet and the electronic ink appears on a regular PC monitor with a resolution of 80 dpi. The default font is Times-12. The resolution of the tablet is 500 dpi.

   The Microsoft editor is used to make corrections. Most subjects find this editor particularly difficult to use. It has a confusing way of automatically switching between writing mode and edit mode. For the default font size and the edit gestures are impossible to use to precisely select, delete, insert, etc. In figure 1 we show a subset of these gestures. It is clear that the position where the edit command should apply cannot be precisely determined with such gestures. Using fewer and simpler gestures (in particular avoiding retracing) is probably a better strategy.

2. Newton cursive recognizer. A recognizer developed by ParaGrah International and distributed by Apple with the Newton personal communicator. It handles mixed style natural handwriting, including cursive It uses a language model. "QWERTY" was added to its dictionary.

   Subjects write on a 3 by 4 inch transparent tablet. The electronic ink is displayed under the pen. The resolution of the display 80 dpi. The font size is about 20 dots.

   A set of second choices of recognition is obtained by double taping a word. A gesture-based editor is used to make corrections. The gestures are found in figure 2. Changes are made by overwriting on the text already transcribed. Most subjects found that the easiest way of making corrections is to rewrite entire words.

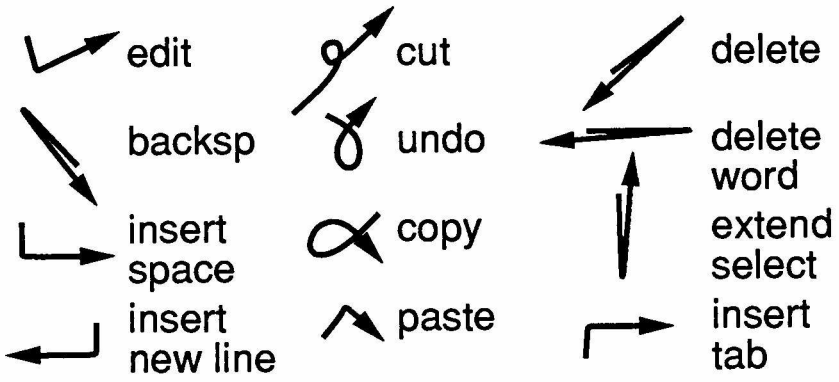

Figure 1: Microsoft edit gestures. The arrow indicates the direction of the stroke.

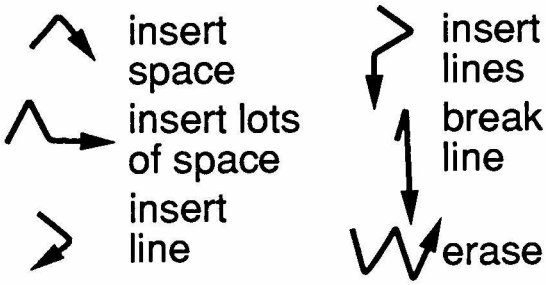

Figure 2: ParaGraph edit gestures. The arrow indicates the direction of the stroke.

3. **Graffiti recognizer.** A recognizer developed and distributed by Palm Computing. It uses a special alphabet (see figure 3). The characters of the alphabet are designed to simplify recognition and segmentation while retaining a resemblance to normal characters. There is a single set of 26 alphabetical characters. Upper and lowercase modes are obtained with a shift gesture. No language model or dictionary is used by the recognizer.

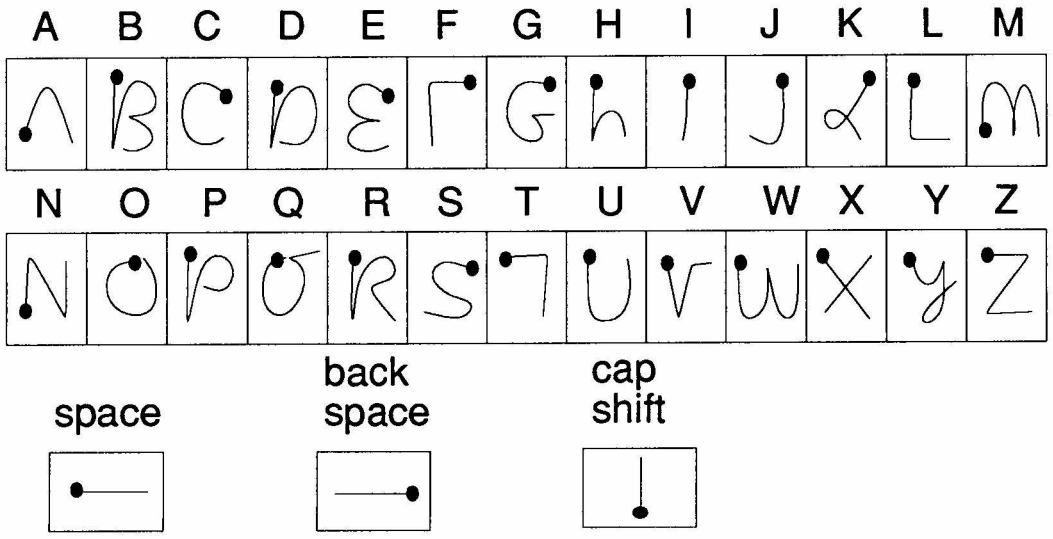

Figure 3: Graffiti alphabet. The dot indicate the starting point.

Subjects write on a 3 by 4 inch transparent tablet (the Tandy personal communicator). The electronic ink is displayed under the pen. The resolution of the display is 80 dpi. The font size is 12 dots. It is possible to avoid moving the hand and to write characters on top of one another.

A gesture-based editor is used to make corrections. Most subjects used backspace to edit and found that it is easier to adjust to a new alphabet than anticipated.

4. Simon predictive keyboard. An interface designed at IBM; it is distributed by Bell South with their Simon cellular phone (see figure 4). It is a 1.5 by 2 inch "software" keyboard with only 6 keys displayed at a time on the touch sensitive screen(figure 4).

   The system uses a language model to predict what is the set of 6 keys which most likely contain the next character. If the desired character is not in that set, a button allows the user to get a new selection of 6 keys. Backspace is used to make corrections. Most users said that finding the position of the right key requires too much concentration.

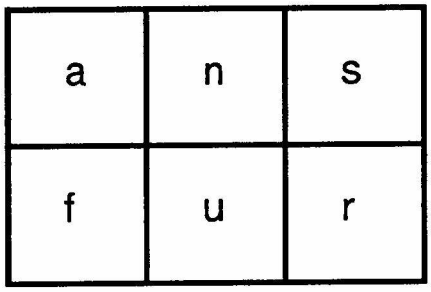

Figure 4: Simon predictive keyboard.

5. Sharp Wizard keyboard. A small keyboard attached to the Sharp Wizard personal organizer. It is a 2 by 4.5 inch "hardware" QWERTY keyboard. Of its normal editing capabilities, only backspace was enabled because the tests were run unattended and teaching the full interface to novice users was too complicated.

6. Mini "software" keyboard. A very small keyboard designed at Bell Laboratories by C. Warwick et al. as an example of cellular phone text entry interface. It is a 1.5 by 2 inch "software" QWERTY keyboard with only uppercase mode. A Dauphin pen computer was used to simulate the cellular phone. The keyboard is displayed on the touch sensitive screen. The display has a resolution of 130 dpi. Users tap on the keys with a stylus. Backspace is used to make corrections.

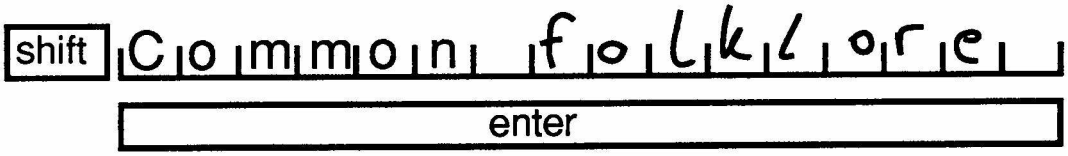

Figure 5: LeRec user interface.

7. LeRec. A handwriting recognizer for handprinted characters entered in a comb (figure 5). It was designed at Bell Laboratories [1]. It is a prototype user interface for small personal communicators. It uses a Sun sparc station connected to a Wacom tablet of resolution 500 dpi with a display of resolution 100 dpi which echoes the electronic ink under the pen. The ink is replaced by the result of recognition. Corrections are made by overwriting. Words are entered with the "enter" key when the recognition is satisfactory. Users can write either uppercase or lowercase. The recognition results always appear in lowercase, unless the shift key has been selected. No language model is used.

8. Sony Magic Link keyboard. A keyboard developed by Genaral Magic as part of the Magic Cap operating system. It is a 2 by 4.5 inch "software" QWERTY keyboard. The keys

Interface number	Stdev grade before (Step A)	Stdev time (sec.) (Step B)	Stdev grade after (Step C)
I	1.01	11.94	1 65
II	1 38	15.27	2 30
5	1 62	50 52	1.30
8	1 35	34 18	1 58
7	1 67	79 40	1.54
3	1 79	215 67	2 21
6	1 42	125.45	1 61
2	2 30	214.95	2.36
4	1 12	89 70	1 69
1	2.73	328 59	2 25

Table 2: Standard deviations for the results of Table 1.

are displayed on the touch sensitive screen of the Sony Magic Link personal digital assistant. The resolution of the display is 106 dpi. The size of the keys is such that, in principle, users could touch type. In practice, most users preferred tapping with a stylus. Backspace is used for editing. Word completion happens ahead of time if the end of the word can be guessed by the language model.

4 Results

The results of the test are shown in tables 1 and 2, ranked in decreasing order of "Grade after".[2] We made the following observations:

- Among the two reference text entry modes, regular keyboard typing (number I) is the favorite, but it is closely followed by natural handwriting on paper (number II). All subjects are computer literate, which may account for this bias in favor of typing.

- For small portable devices which cannot have a regular keyboard, a small "hardware" keyboard (number 5) is the favorite mode of text entry! Even more surprisingly, the Sony Magic Link "software" keyboard (number 8) has performance that are not significantly worse than the small "hardware" keyboard. However, a very small "software" keyboard (number 6) is significantly worse. Both number 5 and number 8 are well accepted interfaces. With LeRec (number 7) they are the only interfaces that get a better "Grade after" than "Grade before". They are only about two times slower than on a regular size keyboard. Expert users claim that it is even possible to touch type on such a keyboard.

[2]The performances of system 7 (LeRec) are calculated on the 13 subjects which tried version 2, which, on average is 34 seconds faster than version 1 (which is already much faster than any of the other recognizers)

The time on system 1 (CIC) takes into account only subjects who did not abandon (8 subjects abandoned because they thought the editor was hopeless)

In general, missing values were not accounted for the average was taken over the values present in the result forms

- The *software* interface which is by far preferred is LeRec (number 7). It is interesting to notice that since the time when we did this study, several interfaces similar to LeRec's interface have appeared on the market (e.g. one from Lexicus and one from Apple). Such handwriting interfaces impose constraints on the user (such as handwriting in boxes or combs and using a shift key to capitalize letters). The resulting increase in handwriting recognition accuracy and the ease of error correction make them a possible candidate to replace small keyboards on devices such as cellular phones and small personal organizers where there is not enough space for a reasonable size keyboard. It is worth noticing however that LeRec is almost 3 times slower than the sharp Wizard keyboard (number 5).

- All other user interfaces have about the same score, except for the CIC interface which is much worse than anything else. The subjects declared that the recognizer was good but that the editor was impossible to use.

- An interesting fact is that keyboard-based interfaces 4 and 6 are faster than handwriting recognition based interfaces 2 and 3 and yet do not get better grades. Otherwise, "Grade after" is correlated with the logarithm of "Time", as shown in figure 6. One can also observe that the standard deviations of 2 and 3 are about twice as large as those of 4 and 6.

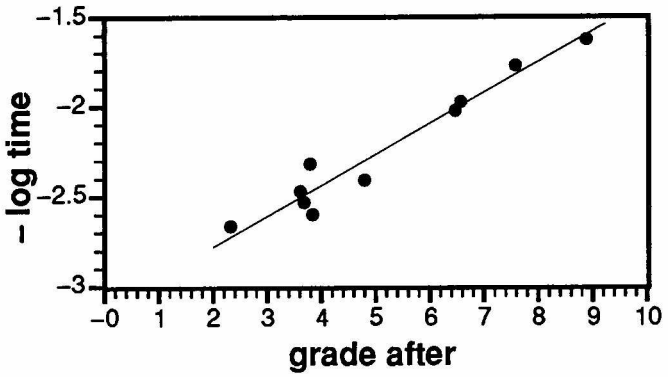

Figure 6: Correlation of "Grade after" and "Time".

These experiments measure "walk-up" usability of text entry interfaces. Expert users report better performance, after struggling with the interface for a few hours. For instance, one user reports peak performance of 96 seconds with Graffiti and 69 seconds with CIC. This user observed that the errors made by the CIC recognizer are easier to correct with a bigger font than the default font which was used for the test. Another user has peak performance of 69 seconds on the Wizard. Yet another one reaches 126 seconds with LeRec.

Humans have such a faculty of adaptation that virtually anything can be made to work with sufficient motivation. Motivation should not be neglected though since many inventors have proposed over the years new and improved keyboard layouts which have all failed to replace the QWERTY keyboard [3]. "The problem is – as it is so often – in changing us" (Bob Lucky - Silicon Dreams [3]).

5 Conclusion and Further Work

From this test, it appears that handwriting has not yet demonstrated its viability as computer user interface. In order to beat the keyboard, everything needs to be done to improve recognition accuracy and simplify the error correction. The user interface designed in our department [1] is a step in this direction. It outperformed all the other handwriting interfaces tested by a wide margin. Our findings are summarized by the graph of Figure 7 which roughly positions the interfaces we tested with respect to one another.

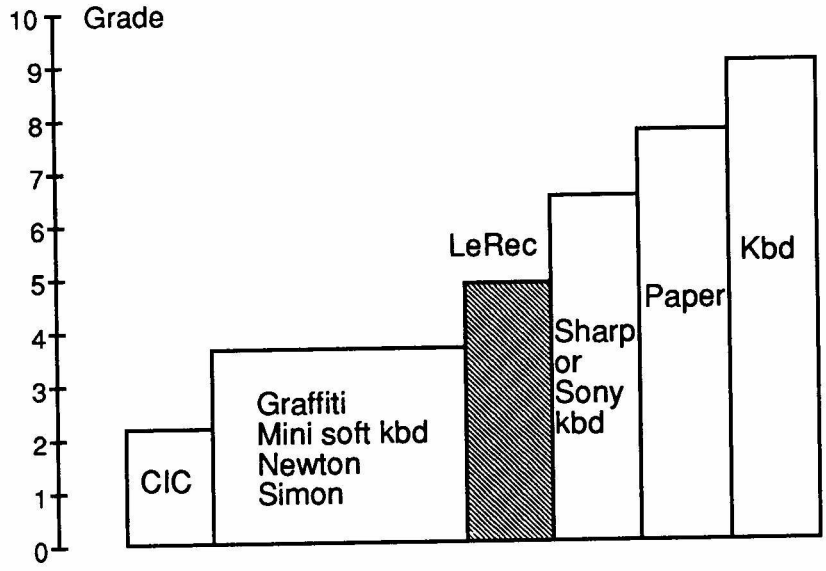

Figure 7: Relative positions of the interfaces tested according to their "grade after".

It has been observed that ease of correction is at least as important as accuracy of text entry and that interfaces cannot be too demanding of user attention to be useful. This should influence the design of future interfaces.

Since the time this study was performed, other interfaces similar to that of LeRec have appear on the market with much greeting from the press. It is too early to say whether they will be welcome by the public.

Acknowledgements

We would like to thank everyone who participated to the user interface test, Corinna Cortes, Yann Le Cun, Vladimir Vapnik, Yoshua Bengio, Urs Müller, Donnie Henderson, Hans-Peter Graf, Larry Jackel, Edi Säckinger, Jane Bromley, Patrice Simard, David Meltzer, Craig Nohl, Sandi von Pier, Isabelle Guyon, John Denker, Eric Cosatto, Markus Schenkel, Matthew Partridge, Michael Joyce, Ping-Wen Ong, and Colin Warwick. Special thanks to Donnie Henderson, Patrice Simard and Craig Nohl for the help they provided to set up the experiments and to John Denker for suggestions on statistics calculations.

References

[1] Y. Bengio, Y. Le Cun, and D. Henderson. Globally trained handwritten word recognition using spatial representation, space displacement neural networks and hidden markov models. In *Advances in Neural Information Processing Systems 6*. Morgan Kauffmann, 1994.

[2] I. Guyon and C. Warwick. Handwriting as computer interface. Technical Report BL0113590-941012-22, AT&T Bell Laboratories, to appear in Joint EC-US Survey of the State of the Art in Human LanguageTechnology (NSF), 1994.

[3] R. Lucky. *Silicon Dreams*. St. Martin's press, New York, 1989.

[4] L. Stampfli. Evolution of pen computing. *Pen Computing*, November 1994.

[5] C. Warwick. Trends and limits in the talk time of personal communicators. Technical report, Proceedings of the IEEE, April, to appear 1995.