BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 1, Number 2, March 1979

# MATHEMATICAL TYPOGRAPHY

BY DONALD E. KNUTH

Dedicated to George Pólya on his 90th birthday

ABSTRACT. Mathematics books and journals do not look as beautiful as they used to. It is not that their mathematical content is unsatisfactory, rather that the old and well-developed traditions of typesetting have become too expensive. Fortunately, it now appears that mathematics itself can be used to solve this problem.

A first step in the solution is to devise a method for unambiguously specifying mathematical manuscripts in such a way that they can easily be manipulated by machines. Such languages, when properly designed, can be learned quickly by authors and their typists, yet manuscripts in this form will lead directly to high quality plates for the printer with little or no human intervention.

A second step in the solution makes use of classical mathematics to design the shapes of the letters and symbols themselves. It is possible to give a rigorous definition of the exact shape of the letter "a", for example, in such a way that infinitely many styles (bold, extended, sans-serif, italic, etc.) are obtained from a single definition by changing only a few parameters. When the same is done for the other letters and symbols, we obtain a mathematical definition of type fonts, a definition that can be used on all machines both now and in the future. The main significance of this approach is that new symbols can readily be added in such a way that they are automatically consistent with the old ones.

Of course it is necessary that the mathematically-defined letters be beautiful according to traditional notions of aesthetics. Given a sequence of points in the plane, what is the most pleasing curve that connects them? This question leads to interesting mathematics, and one solution based on a novel family of spline curves has produced excellent fonts of type in the author's preliminary experiments. We may conclude that a mathematical approach to the design of alphabets does not eliminate the artists who have been doing the job for so many years; on the contrary, it gives them an exciting new medium to work with.

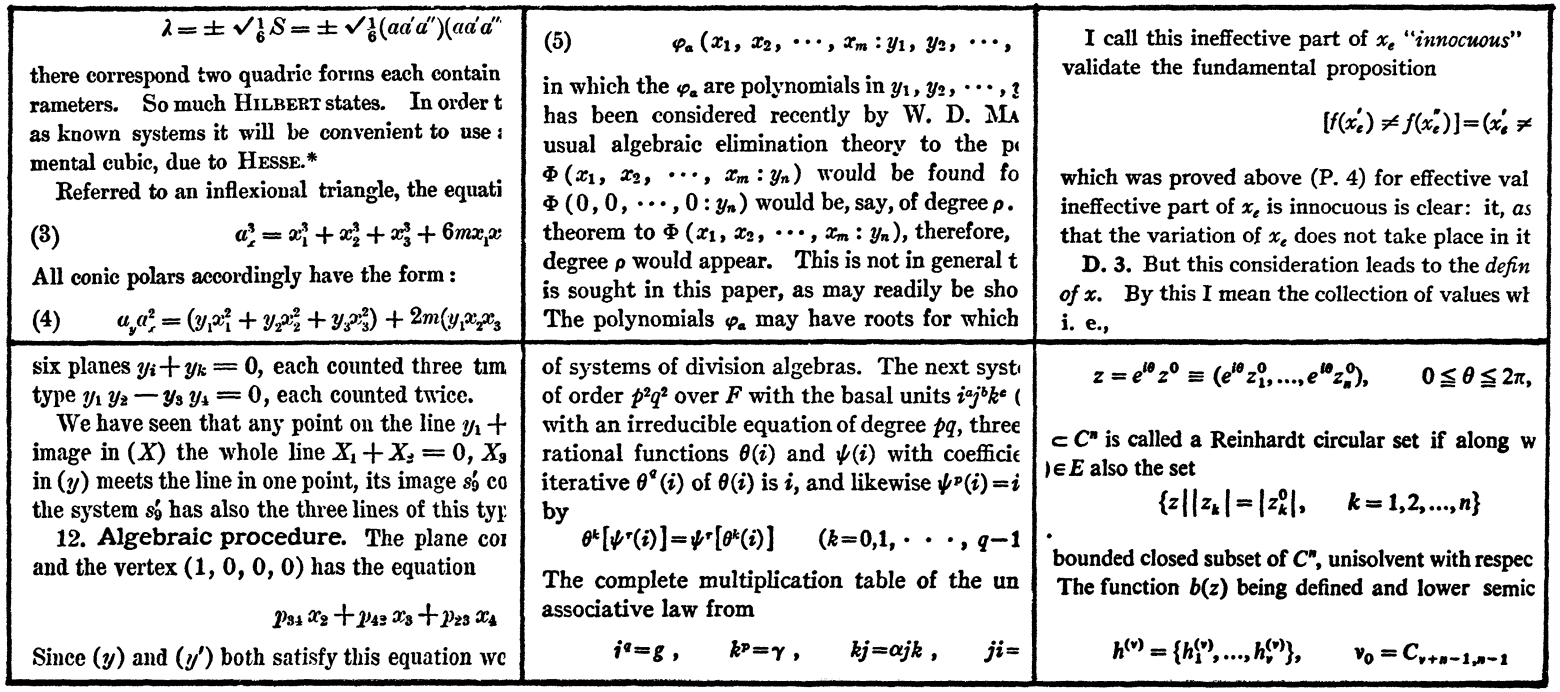
I will be speaking today about work in progress, instead of completed research; this was not my original intention when I chose the subject of this lecture, but the fact is I couldn't get my computer programs working in time. Fortunately it is just as well that I don't have a finished product to describe to you today, because research in mathematics is generally much more interesting while you're doing it than after it's all done. I will try therefore to convey

---

Josiah Willard Gibbs Lecture, given under the auspices of the American Mathematical Society, January 4, 1978; received by the editors February 10, 1978.

The research was supported in part by National Science Foundation grant MCS72-03752 A03, and by the Office of Naval Research contract N00014-76-C-0330. Reproduction in whole or in part is permitted for any purpose of the United States Government.


0002-9904/79/0000-0117/$10.00



$\lambda = \pm \sqrt{\frac{1}{6}} S = \pm \sqrt{\frac{1}{6}} (aa'a'')(aa'a''$

there correspond two quadric forms each contain
rameters. So much HILBERT states. In order t
as known systems it will be convenient to use
mental cubic, due to HESSE.*

Referred to an inflexional triangle, the equati

(3) $a_x^3 = x_1^3 + x_2^3 + x_3^3 + 6mx_1x$

All conic polars accordingly have the form:

(4) $a_y a_x^2 = (y_1x_1^2 + y_2x_2^2 + y_3x_3^2) + 2m(y_1x_2x_3$

six planes $y_i + y_k = 0$, each counted three tim
type $y_1 y_2 - y_3 y_4 = 0$, each counted twice.

We have seen that any point on the line $y_1 +$
image in $(X)$ the whole line $X_1 + X_2 = 0$, $X_3$
in $(y)$ meets the line in one point, its image $s_9'$ co
the system $s_9'$ has also the three lines of this typ

12. **Algebraic procedure.** The plane co
and the vertex (1, 0, 0, 0) has the equation

$p_{34} x_2 + p_{42} x_3 + p_{23} x_4$

Since $(y)$ and $(y')$ both satisfy this equation we

(5) $\varphi_a (x_1, x_2, \cdots, x_m : y_1, y_2, \cdots,$

in which the $\varphi_a$ are polynomials in $y_1, y_2, \cdots,$
has been considered recently by W. D. MA
usual algebraic elimination theory to the p
$\Phi(x_1, x_2, \cdots, x_m : y_n)$ would be found fo
$\Phi(0, 0, \cdots, 0 : y_n)$ would be, say, of degree $\rho$.
theorem to $\Phi(x_1, x_2, \cdots, x_m : y_n)$, therefore,
degree $\rho$ would appear. This is not in general t
is sought in this paper, as may readily be sho
The polynomials $\varphi_a$ may have roots for which

of systems of division algebras. The next syst
of order $p^2q^2$ over $F$ with the basal units $i^a j^b k^c$ (
with an irreducible equation of degree $pq$, three
rational functions $\theta(i)$ and $\psi(i)$ with coefficie
iterative $\theta^q(i)$ of $\theta(i)$ is $i$, and likewise $\psi^p(i) = i$
by

$\theta^k[\psi^r(i)] = \psi^r[\theta^k(i)] \qquad (k = 0,1, \cdots, q-1$

The complete multiplication table of the un
associative law from

$i^q = g, \qquad k^p = \gamma, \qquad kj = \alpha jk, \qquad ji =$

I call this ineffective part of $x_e$ "*innocuous*"
validate the fundamental proposition

$[f(x_e') \neq f(x_e'')] = (x_e' \neq$

which was proved above (P. 4) for effective val
ineffective part of $x_e$ is innocuous is clear: it, *as*
that the variation of $x_e$ does not take place in it

**D. 3.** But this consideration leads to the *defin*
*of x.* By this I mean the collection of values wh
i. e.,

$z = e^{i\theta} z^0 \equiv (e^{i\theta} z_1^0, \ldots, e^{i\theta} z_n^0), \qquad 0 \leqq \theta \leqq 2\pi,$

$\subset C^n$ is called a Reinhardt circular set if along w
$) \in E$ also the set

$\{z \mid |z_k| = |z_k^0|, \qquad k = 1,2,\ldots,n\}$

bounded closed subset of $C^n$, unisolvent with respec
The function $b(z)$ being defined and lower semic

$h^{(\nu)} = \{h_1^{(\nu)}, \ldots, h_\nu^{(\nu)}\}, \qquad \nu_0 = C_{\nu+n-1,n-1}$

FIGURE 1. A sequence of typographical styles in the AMS *Transactions*:
(a) vol **1** (1900), p. 2; (b) vol **13** (1912), p. 135; (c) vol **23** (1922), p. 216;
(d) vol **25** (1923), p. 10; (e) vol **28** (1926), p. 207; (f) vol **105** (1962), p. 340;

$$0 = r_k x(\sum r_i\alpha_i) - (\sum r_i\alpha_i)xr_k = \sum_{i=1}^{k-1}$$

This element is of lower length. It follows there $i = 1, \cdots, k$. Hence, (a) yields that $r_i = \lambda_i r_k$, $\lambda_i$ Now $r_k \neq 0$, by the minimality of $k$, and $\sum\lambda_i$ which we deduce that $\sum\lambda_i\alpha_i = 0$. But the $\alpha_i$ a which is impossible since in particular $\lambda_k = 1$.

THEOREM 7. *Let $R$ be a dense ring of linear t. $F$ be a maximal commutative subfield $D$. If $R_F$ tion of finite rank over $F$, then $R$ contains als*

The set $N_1$ is nowhere dense in $Z_1$ and thus $N = \rho$

For each $\zeta \in Y - N$ we must prove that $f_\zeta$ satisfi be the unique projection in $\{P_d \mid d \in D\}$ such that the algebra $(E\mathscr{A}E)\cdot P_0$ is finite and homogeneou onal abelian projections $E_1, E_2, \ldots, E_n$ such that $(1 \leqq j, k \leqq n)$ be partial isometric operators in $(E\mathscr{A}$

(1) $U_{jk}U_{lm} = \delta_{mj}U_{lk}$, where $\delta$ is the Kronecker d
(2) $U_{jk}^* = U_{kj}$; and
(3) $U_{jj} = E_j$,

for all $1 \leqq j, k, l, m \leqq n$. For each $A$ in $(E\mathscr{A}E)\cdot P_0$, t in $\mathscr{Z}_1 P_0$ such that

*The algebra $P$ is nearly simple if and only if th*

(a) *$N$ is spanned by $a, \cdots, a^{n-k-1}, b_1, \cdots$ $i, j = 1, \cdots, k$.*

(b) *Either $n - k =$ char $F$ with $k$ even or $n$*

Proof. By Theorem 5.5, there are elements $a, \cdots, a^{n-k-1}, b_1, \cdots, b_k$. Furthermore, $ab_i =$ for all $i, j$ where each $\alpha_i, \lambda_{ij}$ is in $F$. From t space of the space spanned by $a^{n-k-1}, b_1, \cdots,$

Assume $P$ is nearly simple. Then there is show that each $b_i$ is in $M$. To do this, it is nec

:tions in GL($W$) and $h_{\alpha\beta}$, $\alpha, \beta \in I$ as coordinate ined by the respective bases chosen above. If $\alpha$, nction of $\wedge^p$ is the minor of $|g_{ij}|$ determined by columns $\beta(1), \ldots, \beta(p)$. The coordinate ring of $_{\alpha\beta}$ together with $1/\det|h_{\alpha\beta}|$, while that of GL($W$) ogether with $1/\det|g_{ij}|$. The coordinate functions , so to show $\wedge^p$ is a morphism it suffices to show nial in $g_{ij}$ and $1/\det|g_{ij}|$. For this, the following

*haracter of* GL($W$) *is an integral power of the*

of $Q$, i.e.

$i) = 0$ for every $x \in A$ for which $x(Q) = 0\}$.

'or $m_A$ is equivalent to the one induced by the :

$\{|x(z)|: x \in A, \|x\| \leqslant 1$ and $x(w) = 0\}$.

:present the open unit disk in the complex plane, $\mathbf{C}$, t polydisk in $n$-dimensional complex space $\mathbf{C}^n$. $T^n$ oundary of $D^n$, i.e.

:onverges pathwise to $X^\lambda$, and uniformly for $t \in$ for which $X_i$ is the (last) minimum of $Y^\lambda$, let $Y_i^\lambda$, ılues of $Y^\lambda$, and $T_i^\lambda$ the interjump times for $Y^\lambda$ st an $i$ such that $Y_i^\lambda = T_i^\lambda = \infty$. Notice that $Y_Q^\lambda$ i ıd that as $\varepsilon \to 0$, $Y_Q^\lambda$ converges to $I^\lambda = \inf_s X_s^\lambda$. Le ts of $(-\infty, \infty)$. Then, for example, if $i \geqslant 1$

$-i \in B, Y_{Q+k}^\lambda - Y_Q^\lambda \in C, T_{Q+k}^\lambda \in D, N > Q > i\}$

$\in A, T_{l-i}^\lambda \in B, Y_{l+k}^\lambda - Y_l^\lambda \in C, T_{l+k}^\lambda \in D, N > Q$

FIGURE 1 [continued]:

(g) vol **114** (1965), p. 216; (h) vol **125** (1966), p. 38; (i) vol **169** (1972), p. 232;
(j) vol **179** (1973), p. 314; (k) vol **199** (1974), p. 370; (l) vol **225** (1977), p. 372.

in this lecture why I am so excited about the project on which I am currently working.

My talk will be in two parts, based on two different meanings of its title. First I will speak about mathematical typography in the sense of typography as the servant of mathematics: the goal here is to communicate mathematics effectively by making it possible to publish mathematical papers and books of high quality, without excessive cost. Then I will speak about mathematical typography in the sense of mathematics as the servant of typography: in this case we will see that mathematical ideas can make advances in the art of printing.

**Preliminary examples.** To set the stage for this discussion I would like to show you some examples by which you can "educate your eyes" to see mathematics as a printer might see it. These examples are taken from the *Transactions of the American Mathematical Society*, which began publication in 1900; by now over 230 volumes have been published. I took these volumes from the library shelves and divided them into equivalence classes based on what I could perceive to be different styles of printing: two volumes were placed into the same class if and only if they appeared to be printed in the same style. It turns out that twelve different styles can be distinguished, and it will be helpful for us to look at them briefly.

The first example (Figure 1a) comes from p. 2 of *Transactions* volume 1; I have shown only a small part of the page in order to encourage you to look at the individual letters and their positions rather than to read the mathematics. This typeface has an old-fashioned appearance, primarily because the upper case letters and the taller lower case ones like '$h$' and '$k$' are nearly twice as tall as the other lower case letters, and this is rarely seen nowadays. Notice the style of the italic letter '$x$', the two strokes having a common segment in the middle. The subscripts and superscripts are set in rather small type.

This style was used in volumes 1 to 12 of the *Transactions*, and also in the first 21 pages of volume 13. Then page 22 of volume 13 introduced a more modern typeface (Figure 1b). In this example the subscripts are still in a very small font, and unfortunately the Greek $\alpha$ here is almost indistinguishable from an italic '$a$'. Notice also that the printer has inserted more space before and after parentheses than we are now accustomed to. During the next few years the spacing within formulas evolved gradually but the typefaces remained essentially the same up through volume 24: with one exception.

The exception was volume 23 in 1922 (Figure 1c), which in my opinion has the most pleasing appearance of all the *Transactions* volumes. This modern typeface is less condensed, making it more pleasant to read. The italic letters have changed in style too, not quite so happily—note the '$x$', for example, which is not as nice as before—but by and large one has a favorable impression when paging through this volume. Such quality was not without its cost, however; according to a contemporary report in the AMS *Bulletin* [**45**, p. 100], the *Transactions* came out 18 months late at the time! Perhaps this is why the Society decided to seek yet another printer.

In order to appreciate the next change, let's look quickly at two excerpts from the *Bulletin* relating to the very first Gibbs lecture (Figure 2). The

## THE JOSIAH WILLARD GIBBS LECTURESHIP

The Council of the Society has sanctioned the establishment of an honorary lectureship to be known as the Josiah Willard Gibbs Lectureship. The lectures are to be of a popular nature on topics in mathematics or its applications, and are to be given by invitation under the auspices of the Society. They will be held annually or at such intervals as the Council may direct. It is expected that the first lecture will be delivered in New York City during the winter of 1923–24, and a committee has been authorized to inaugurate the lectures by choosing the first speaker and making the necessary arrangements.

R. G. D. RICHARDSON,
*Secretary.*

## THE FIRST JOSIAH WILLARD GIBBS LECTURE

The first Josiah Willard Gibbs Lecture was delivered under the auspices of this Society on February 29, 1924, by Professor M. I. Pupin, of Columbia University, in the auditorium of the Engineering Societies' Building, New York City. A large and distinguished audience was present, including, besides members of the Society, many physicists, chemists, and engineers who had been invited to attend.

In introducing the speaker, President Veblen spoke as follows:

"In instituting the Willard Gibbs Lectures, the American Mathematical Society has recognized the dual character of mathematics. On the one hand, mathematics is one of the essential emanations of the human spirit,—a thing to be valued in and for itself, like art or poetry. Gibbs made

FIGURE 2. A time of transition.
(Excerpts from the AMS Bulletin **29** (1923), p. 385; **30** (1924), p. 289.)

preliminary announcement in 1923 appeared in the modern typeface used during that year, but the letter shapes in the report of the first lecture in 1924 were very cramped and stilted. The upper case letters in the title are about the same, but the lower case letters in the text are completely different.

This same style appeared in volume 25 of the *Transactions* (Figure 1d), which incidentally was set in Germany in order to reduce the cost of printing. Note that the boldface letters and the italic letters in this example are actually quite beautiful—and we're back to the good old style of '$x$' again—so the mathematical formulas looked great while the accompanying text was crowded. Fortunately only three volumes were published in this style.

A new era for the *Transactions* began in 1926, when its printing was taken over by the Collegiate Press in Menasha, Wisconsin. Volumes 28 through 104 were all done in the same style, covering 36 years from 1926 to 1961, inclusive, and this style (Figure 1e) was used also in the *American Mathematical Monthly*. In general the modern typefaces were quite satisfactory, but there was also a curious anomaly: Italic letters used in subscripts and superscripts of mathematical formulas were in a different style from those used on the main line! For example, notice the two $k$'s in the first displayed formula of Figure 1e: the larger one has a loop, so it is topologically different from the smaller one. Similarly you can see that the $p$ in $k^p$ is quite different from the $p$ in $p^2$. There are no $x$'s in this example, but if you look at other pages you will find that the style of $x$ that I like best appears only in subscripts and superscripts. I can't understand why this discrepancy was allowed to persist for so many years.

Another period of typographic turmoil for the *Transactions* began with volume 105 in 1962. This volume, which was typeset in Israel, introduced a switch to the Times Roman typeface (Figure 1f); an easy way to recognize the difference quickly is to look at the shading on the letter "o", since it now is somewhat slanted; in the previously used fonts this letter always was more symmetrical, as if it were drawn with a pen held horizontally, but in Times Roman it clearly has an oblique stress as if it were drawn by a right-handed penman. Note that the three $k$'s are topologically the same in the displayed equation here; but for some reason the two subscript $k$'s are of different sizes. Many of the Times Italic letters have a somewhat different style than readers of the *Transactions* had been accustomed to, and I personally think that this font tends to make formulas look more crowded. Actually the changeover to Times Roman and Times Italic wasn't complete; the italic letter $g$ still had its familiar shape, perhaps because the new shape looked too strange to mathematicians.

Volumes 105 through 124 were all done in this style, except for a brief interruption: In volumes 114, 115, and 116 the shading on the $o$'s was symmetrical and the $k$'s had loops (Figure 1g). Another style was used for volumes 125–168 (Figure 1h): again Times Roman was the rule, even in the $g$'s, except for subscripts and superscripts which were in the style I prefer; for example, compare the $j$'s and $k$'s. (These latter volumes were typeset in Great Britain.)

A greatly increased volume of publication, together with the rising salaries of skilled personnel, was making it prohibitively expensive to use traditional

methods of typesetting, and the Society eventually had to resort to a fancy form of typewriter composition that could simply be photographed for printing. This unfortunate circumstance made volumes 169–198 of the *Transactions* look like Figure 1i, except for volumes 179, 185, 189, 192, 194, and 198, which were done in a far better (yet not wholly satisfactory) style that can be distinguished from Figure 1f by the italic *g*'s. Figure 1j was composed on a computer using a system developed by Lowell Hawkinson and Richard McQuillin; this was one of the fruits of an AMS research project supported by the National Science Foundation [2], [3], [4], [5], [6].

Computer typesetting of mathematics was still somewhat premature at the time, however, and another kind of "cold copy" made its appearance in volumes 199 through 224—an "IBM Compositor" was used, except for volumes 208 and 211 which reverted to the Varityper style of Figure 1i. The new alphabet was rather cramped in appearance, and some words were even more crowded than the others (see Figure 1k). At this point I regretfully stopped submitting papers to the American Mathematical Society, since the finished product was just too painful for me to look at. Similar fluctuations of typographical quality have appeared recently in all technical fields, especially in physics where the situation has gotten even worse. (The history of publication at the American Society of Civil Engineers has been discussed in an interesting and informative article by Paul A. Parisi [44].)

Fortunately things are now improving. Beginning with volume 225, which was published last year, the *Transactions* now looks like Figure 1l; like Figure 1j, it is computer composed, and the Times Roman typeface is now somewhat larger. I still don't care for this particular style of italic letters, and there are some bugs needing to be ironed out such as the overlap between lines shown in this example; but it is clear that the situation is getting better, and perhaps some day we will once again be able to approach the quality of volumes 23 and 24.

**Computer-assisted composition.** Perhaps the main reason that the situation is improving is the fact that computers are able to manipulate text and convert it into a form suitable for printing. Experimental systems of this kind have been in use since the early 1960s (cf. the book by Barnett [**10**]), and now they are beginning to come of age. Within another ten years or so, I expect that the typical office typewriter will be replaced by a television screen attached to a keyboard and to a small computer. It will be easy to make changes to a manuscript, to replace all occurrences of one phrase by another and so on, and to transmit the manuscript either to the television screen, or to a printing device, or to another computer. Such systems are already in use by most newspapers, and new experimental systems for business offices actually will display the text in a variety of fonts [**26**]. It won't be long before these machines change the traditional methods of manuscript preparation in universities and technical laboratories.

Mathematical typesetting adds an extra level of complication, of course. Printers refer to mathematics as "penalty copy", and one of America's foremost typographers T. L. De Vinne wrote that "[even] under the most favorable conditions algebra will be troublesome." [**17**, p. 171.] The problem

| Formula | Type C | Type B | Type T |
|---|---|---|---|
| $\frac{1}{2}$ | `$f1$s2$t` | `1 over 2` | `1 \over 2` |
| $\theta^2$ | `*gq"2` | `theta sup 2` | `\theta↑2` |
| $\sqrt{f(x_i)}$ | `$rf(x'i)$t` | `sqrt{f(x sub i )}` | `\sqrt{f(x↓i)}` |

FIGURE 3. Three ways to describe a formula.

used to be that the two-dimensional formulas required complicated positioning of individual metal pieces of type; but now this problem reduces to a much simpler one, namely that two-dimensional formulas need to be represented as a one-dimensional sequence of instructions for transmission to the computer.

One-dimensional languages for mathematical formulas are now familiar in programming languages such as FORTRAN, but a somewhat different approach is needed when all of the complexities of typesetting are considered. In order to show you the flavor of languages for mathematical typesetting, I will briefly describe the three reasonably successful systems known to me. The first, which I will call Type C, is typical of the commercially available systems now used to typeset mathematical journals (cf. [12]). The second, which I will call Type B, was developed at Bell Telephone Laboratories and has been used to prepare several books and articles including the article that introduced the system [27]. The third, which I will call Type T, is the one I am presently developing as part of the system I call TEX [29].[1]

Figure 3 shows how three simple formulas would be expressed in these three languages. The Type C language uses \$f . . . \$s . . . \$t for fractions, *$g$ for "the next character is Greek", $q$ for the Greek letter theta, $''$ for superscripts, \$r . . . \$t for square roots and $'$ for subscripts. The Type B language is more mnemonic, using "over", "theta", "sup", "sqrt", and "sub" together with braces for grouping when necessary. The Type T language is similar but it does not make use of "reserved words"; a special character \ is used before any nonstandard text. This means that spaces can be ignored, while they need to be inserted in just the right places in the Type B language; for example, the space after the "$i$" is important in the example shown, otherwise $f(x_i)$ would become $f(x_{i)}$ according to the Type B rules. Another reason for the \ delimiter in Type T is that it becomes unnecessary to match each text item against a stored dictionary, and it is possible to use "sup" to mean supremum instead of superscript. The special symbols \ { } ↑ ↓ in Type T can be changed to any other characters if desired; although these five

[1]This has no connection with a similarly-named system recently announced by Honeywell Information Systems, or with another one developed by Digital Research. In my language, the T, E, and X are Greek letters and TEX is pronounced "tech", following the Greek words for art and technology.

symbols don't appear on conventional typewriters, they are common on computer terminal keyboards.

Incidentally, computer typesetting brings us some good news: It is now quite easy to represent square roots in the traditional manner with radical signs and vincula, so we won't have to write $x^{1/2}$ when we don't want to.[2]

None of these languages makes it possible to *read* complex formulas as easily as in the two-dimensional form, but experience shows that it is not difficult for untrained personnel to learn how to type them. According to [12], "Within a few hours (a few days at most) a typist with no math or typesetting background can be taught to input even the most complex equations." And the Type B authors [27] report that "the learning time is short. A few minutes gives the general flavor, and typing a page or two of a paper generally uncovers most of the misconceptions about how it works." Thus it will be feasible for both typists and mathematicians to prepare papers in such a language, without investing a great deal of effort in learning the system. The only real difficulties arise when preparing tables that involve tricky alignments.

Once such systems become widespread, authors will be able to prepare their papers and see exactly how they will look when printed. Everyone who writes mathematical papers knows that his intentions are often misunderstood by the printer, and corrections to the galley proofs have a nontrivial probability of introducing further errors. Thus, in the words of three early users of the Bell Labs' system, "the moral seems clear. If you let others do your typesetting, then there will be errors beyond your control; if you do your own, then you have only yourself to blame." [1] Personally, I can't adequately describe how wonderful it feels when I now make a change to the manuscript of my book, as it is stored in the Stanford computer, since I *know* that the change is immediately in effect; it never will go through any middlemen who might misunderstand my intention.

Perhaps some day a typesetting language will become standardized to the point where papers can be submitted to the American Mathematical Society from computer to computer via telephone lines. Galley proofs will not be necessary, but referees and/or copy editors could send suggested changes to the author, and he could insert these into the manuscript, again via telephone.

Of course I am hoping that if any language becomes standard it will be my TEX language. Well . . . perhaps I am biased, and I know that TEX provides only small refinements over what is available in other systems. Yet several dozen small refinements add up to something that is important to me, and I think such refinements might prove important to other people as well. Therefore I'd like to spend the next few minutes explaining more about TEX.

**The TEX input language.** TEX must deal with "ordinary" text as well as mathematics, and it is designed as a unified system in which the mathematical features blend in with the word-processing routines instead of being "tacked on" to a conventional typesetting language. The main idea of TEX is to

[2](ADDED IN PROOF). I was pleased to find that this announcement was greeted with an enthusiastic round of applause when I delivered the lecture.

construct what I call *boxes*. A character of type by itself is a box, as is a solid black rectangle; and we use such "atoms" to construct more complex boxes analogous to "molecules", by forming horizontal or vertical lists of boxes. The final pages of text are boxes made out of lists of boxes made out of lists of boxes, and so on down to the individual characters and black rectangles, which are not decomposed further. For example, a typical page of a book is a box formed from vertical lists of boxes representing lines of type, and these lines of type are boxes formed from a horizontal list of boxes representing individual letters. A mathematical formula breaks down into boxes in a natural way; for example, the numerator and denominator of a fraction are boxes, and so is the bar line between them (since it is a thin but solid black rectangle). The elements of a rectangular matrix are boxes, and so on.

The individual boxes of a horizontal list or a vertical list are separated by a special kind of elastic mortar which I call "glue". The glue between two boxes has three component parts $(x, y, z)$ expressed in units of length:

the *space* component, $x$, is the ideal or normal space desired between these boxes;

the *stretch* component, $y$, is the amount of extra space that is tolerable;

the *shrink* component, $z$, is the amount of space that may be removed if necessary.

Suppose the list contains $n + 1$ boxes $B_0, B_1, \ldots, B_n$ separated by $n$ globs of glue having specifications $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$. When this list is made into a box, we *set the glue* according to the desired final size of the box. If the final size is to be larger than we would obtain with the normal spacing $x_1 + \cdots + x_n$, we increase the space proportional to the $y$'s so that the actual space between boxes is

$$x_1 + ty_1, \ldots, x_n + ty_n$$

for some appropriate $t > 0$. On the other hand if the desired final size must be smaller, we decrease the space to

$$x_1 - tz_1, \ldots, x_n - tz_n,$$

in proportion to the individual shrinkages $z_i$. In the latter case $t$ is not allowed to become greater than 1; the glue will never be smaller than $x - z$, although it might occasionally become greater than $x + y$. Once the glue has been set, the box is rigid and never changes its size again.

Consider, for example, a normal line of text, which is a list of individual character boxes. The glue between letters of a word will have $x = y = z = 0$, say, meaning that this word always has the letters butting against each other; but the glue between words might have $x$ equal to the width of the letter '*e*', and $y = x$, and $z = \frac{1}{2}x$, meaning that the space between words might expand or shrink. The spaces after punctuation marks like periods and commas might be allowed to stretch at a faster rate but to shrink more slowly.

An important special case of this glue concept occurs when we have "infinite" stretchability. Suppose the $x$ and $z$ components are zero, but the $y$ component is extremely large, say $y$ is one mile long. If such an element of glue is placed at the left of a list of boxes, the effect will be to put essentially

all of the expansion at the left, therefore the boxes will be right-justified so that the right edge will be flush with the margin. Similarly if we place such infinitely stretchable glue at both ends of the list, the effect will be to center the line. These common typographic operations therefore turn out to be simple special cases of the general idea of variable glue, and the computer can do its job more elegantly since it is dealing with fewer primitives. Incidentally you will notice from this example that glue is allowed to appear at the ends of a list, not just between boxes; actually it is also possible to have glue next to glue, and boxes next to boxes, so that a list of boxes really is a list of boxes and glue mixed in any fashion whatever. I didn't mention this before, because for some reason it seems easier to explain the idea first in the case when boxes alternate with glue.

The same principles apply to vertical lists. For example, the glue which appears above and below a displayed equation will tend to be stretchable and shrinkable, but the glue between lines of text will be calculated so that adjacent base lines will be uniformly spaced when possible. You can imagine how the concept of glue allows you to do special tricks like backspacing (by letting $x$ be negative), in a natural manner.

**Line division.** One of the more interesting things a system like TEX has to do is to divide up a paragraph into individual lines so that each line is about the right length. The traditional way to do this, which is still used on today's computer typesetting systems, is to make the best possible line division you can whenever you come to the right margin, but once this line has been output you never reconsider it—you start the next line with no memory of what has come before. Actually it often happens that one could do better by moving a short word down from one line to the next, but the problem is that you don't know what the rest of the paragraph will be like when you have only looked at one line's worth.

The TEX system will introduce a new approach to the problem of line division, in which the end of a paragraph *does* influence the way the first lines are broken; this will result in more even spacing and fewer hyphenated words. Here is how it works: First we convert the line division problem to a precisely-defined mathematical problem by using TEX's glue to introduce the concept of "badness". When a horizontal list of boxes has a certain natural width $w$ (based on the width of its boxes and the space components of its glue), a certain stretchability $y$ (the sum of the stretch components) and a certain shrinkability $z$ (the sum of the shrinkages), the *badness* of setting the glue to make a box of width $W$ is defined to be $1 + 100t^3$ in our previous notation; more precisely, it is

$$\begin{cases} 1, & \text{if } W = w, \\ 1 + 100\left(\dfrac{W - w}{y}\right)^3, & \text{if } W > w, \\ 1 + 100\left(\dfrac{w - W}{z}\right)^3, & \text{if } w - z \leqslant W < w, \\ \text{infinite}, & \text{if } W < w - z. \end{cases}$$

Thus if the desired width $W$ is near the natural width $w$, or if there is a lot of stretchability and shrinkability, the badness rating is very small; but if $W$ is much greater than $w$ and there isn't much ability to stretch, we have a lot of badness. Furthermore we add *penalty points* to the badness rating if the line ends at a comparatively undesirable place; for example, when a word needs to be hyphenated, the badness goes up by 50, and an even worse penalty is paid if we have to break up mathematical formulas.

The line division problem may now be stated as follows. "Given the text of a paragraph and the set of all allowable places to break it between lines, find breakpoints which minimize the sum of the squares of the badnesses of the resulting lines." This definition is quite arbitrary, of course, but it seems to work. Preliminary experiments show that the same choice of breakpoints is almost always found when simply minimizing the sum of the individual badnesses rather than the sum of their squares, but it seems wise to minimize the sum of squares as a precautionary measure since this will also tend to minimize the maximum badness.

Just stating the line division problem in mathematical terms doesn't solve it, of course; we need to have a good way to find the desired breakpoints. If there are $n$ permissible places to break (including all spaces between words and all possible hypenations), there are $2^n$ possible ways to divide up the paragraph, and we would never have time to look at them all. Fortunately there is a technique that can be used to reduce the number of computational steps to order $n^2$ instead of $2^n$; this is a special case of what Richard Bellman calls "dynamic programming." Let $f(j)$ be the minimum sum of badness squares for all ways to divide the initial text of the paragraph up to breakpoint $j$, including a break at $j$, and let $b(i,j)$ be the badness of a line that runs from breakpoint $i$ to breakpoint $j$. Let breakpoint 0 denote the beginning of the paragraph; and let breakpoint $n+1$ be the end of the paragraph, with infinitely expandable glue inserted just before this final breakpoint. Then

$$f(0) = 0;$$

$$f(j) = \min_{0 \le i < j} \big(f(i) + b(i,j)^2\big), \quad \text{for } 1 \leqslant j \leqslant n+1.$$

The computation of $f(1), \ldots, f(n+1)$ can be done in order $n^2$ steps, and $f(n+1)$ will be the minimum possible sum of badnesses squared. By remembering the values of $i$ at which the minima occurred for each $j$, we can find breakpoints that give the best line divisions, as desired.

In practice we need not test extremely unlikely breakpoints; for example, there is rarely any reason to hyphenate the very first word of a paragraph. Thus it turns out that this dynamic programming method can be further improved to an algorithm whose running time is almost always of order $n$ instead of $n^2$, and comparatively few hyphenations will need to be tried. Incidentally, the problem of hyphenation itself leads to some interesting mathematical questions, but I don't have time to discuss them today. (Cf. **[41]** and the references in that paper.)

The idea of badness ratings applies in the vertical dimension as well as in the horizontal; in this case we want to avoid breaking columns or pages in a

bad manner. For example, penalty points are given for splitting a paragraph between pages after a hyphenation, or for dividing it in such a way that only one of its lines—a so-called "widow" line—appears on a page. The placement of illustrations, tables, and footnotes is also facilitated by formulating appropriate rules of placement in terms of badness.

There is more to TEX, including for example some facilities for handling the rather intricate layouts often needed to typeset tables without having to calculate column widths; but I think I have described the most important principles of its organization. During the next few months I plan to write the computer programs for TEX in such a way that each algorithm is clearly explained and so that the system can be implemented on many different computers without great difficulty; then I intend to publish the programs in a book so that everyone who wants to can use them.

**Entr'acte.** I said at the beginning that this talk would be in two parts, discussing both the ways that typography can help mathematics and that mathematics can help typography. So far we have seen a little of both, but the mathematics has been comparatively trivial. In the remainder of my lecture I would like to discuss what I believe is a much more significant application of mathematics to typography, namely to the specification of the letter shapes themselves. A more accurate way to describe the two parts of my lecture would be to say that the first part was about TEX, a system which takes manuscripts and converts them into specifications about where to put each character on each page; and the second part will be about another system I'm working on called METAFONT, which generates the characters themselves, for use in the inkier parts of the printing business.

Before I get into the second part of my lecture I need to discuss recent developments in printing technology. The most reliable way to print mathematics books of high quality during the past several decades has been to use the monotype process[3] which casts characters in hot lead, together with hand operations for complex built-up formulas. When I watched this process being applied to my own books several years ago, I was surprised to learn that the lead type was used to print only *one* copy; this master copy was then photographed, and the real printing took place from the photographic plates. This somewhat awkward sequence of steps was justified because it was the best way known to give good results. During the 1960s, however, hot lead type was replaced for many purposes by devices like the Photon machine used to prepare the printed programs for this lecture; in this case the process is entirely photographical, since the letter shapes are stored as small negatives on a rotating disk, and the plates needed for printing are obtained by exposing the film after transforming the characters into the proper size and position with mirrors and lenses (cf. [**10**]). Such machines are limited by slow speed and the difficulties of adding new characters.

**"Third-generation" typesetting equipment.** More recent machines, such as the one used to prepare the current volumes of the *Transactions*, have replaced these "analog" processes by a "digital" one. The new idea is to

[3]Actually the Monotype Corporation now manufactures digital photosetting equipment as well as the traditional 'monotype' machines.

divide the page or the photographic negative into millions of tiny rectangles, like a piece of graph paper or like a television screen but with a much higher resolution of about 1000 lines per inch. For each of the tiny "pixels" in such a raster pattern—there are about a million square pixels in every square inch—the typesetting machine decides whether it is to be black or white, and the black ones are exposed on the photographic plate by using a very precisely controlled electron beam or laser beam. Since these machines have few moving parts and require little or no mechanical motion, they can operate at very high speeds even though they are exposing only a tiny bit of the film at any time.

Stating this another way, the new printing equipment essentially treats each page of a book as a huge matrix of 0's and 1's, with ink to be placed in the positions that are 1 while the 0 positions are to be left blank. It's like the flashcards at a football stadium, although on a much grander scale. The total job of a system like TEX now becomes one of converting an author's manuscript into a gigantic bit matrix.

The first question we must ask, of course, is, "What happens to the quality?" Clearly a television picture is no match for a photograph, and the digital typesetting machines would be quite unsatisfactory if their output looked inferior to the results obtained with metal type. In matters like this, I have to confess being somewhat of a stickler and a perfectionist; for example, I refuse to eat margarine instead of butter, and I have never heard an electronic organ that sounds even remotely as beautiful as a pipe organ. Therefore I was quite skeptical about digital typography, until I saw an actual sample of what was done on a high quality machine and held it under a magnifying glass: It was impossible to tell that the letters were generated with a discrete raster! The reason for this is not that our eyes can't distinguish more than 1000 points per inch; in appropriate circumstances they can. The reason is that particles of *ink* can't distinguish such fine details—you can't print the edge of an ink line that zigzags 1000 times on the diagonal of a square inch, the ink will round off the edges. In fact the critical number seems to be more like 500 than 1000. Thus the physical properties of ink cause it to appear as if there were no raster at all.

It now seems clear that discrete raster-based printing devices will soon make the other machines obsolete for nearly all publishing activity. Thus in future days the fact that Gutenberg and others invented movable type will not be especially relevant; it will merely be a curious historical fact that influenced history for only about 500 years. The ultimately relevant thing will be mathematics: the mathematics of matrices of 0's and 1's!

**Semiphilosophical remarks.** I have to tell the next part of the story from my personal point of view. As a combinatorial mathematician, I really identify with matrices of 0's and 1's, so when I learned last spring about such printing machines it was impossible for me to continue what I was doing; I just had to take time off to explore the possibilities of the new equipment. My motivation was also increased by the degradation of quality I had been observing in technical journals; and furthermore the publishers of my books on computer programming had tried valiantly but unsuccessfully to produce the second edition of volume 2 in the style of the first edition without using the

rapidly-disappearing hot lead process. It appeared that my books would soon have to look as bad as the journals! When I saw that these problems could all be solved by appropriate computer programming, I couldn't resist trying to find a solution by myself.

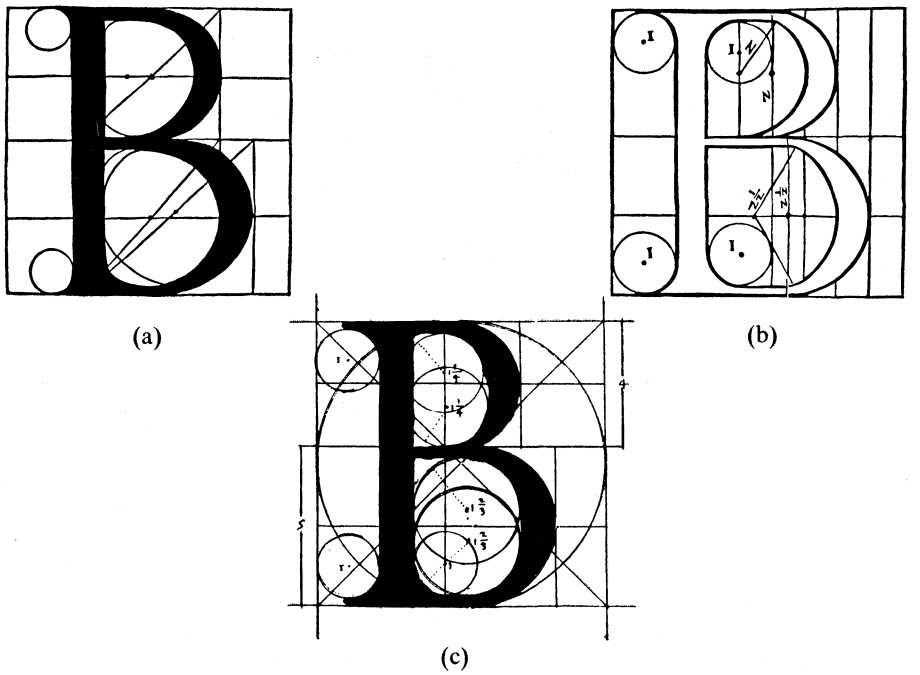
One of the most important factors in my motivation was the knowledge that the problem would be solved once and for all, if I could find a purely mathematical way to define the letter shapes and convert them to discrete raster patterns. Even though new printing methods are bound to be devised in the future, possibly even before I finish volume seven of the books I'm writing, any new machines are almost certain to be based on a high precision raster; and although the precision of the raster may change, the letter shapes can stay the same forever, once they are defined in a machine-independent form. My goal was therefore to give a precise description of the shapes of all the symbols I would need.

I looked at the way fonts of type are being digitized at several places in different parts of the world; it is basically done by taking existing fonts and copying them using sophisticated camera equipment and computer programs, together with manual editing. But this seemed instinctively wrong to me, partly because the sophisticated equipment wasn't readily available in our laboratory at Stanford, and partly because the copying of copyrighted fonts is of questionable legality, but mostly because I felt that the whole idea of making a copy was not penetrating to the heart of the problem. It reminded me of the anecdote I had once heard about slide rules in Japan. According to this story, the first slide rule ever brought to the Orient had a black speck of dirt on it; so for many years all Japanese slide rules had a useless black spot in this same position! The story is probably apocryphal, but the point is that we should copy the substance rather than the form. I felt that the right question to ask would not be "How should this font of type be copied?" but rather: "If the great type designers of the past were alive today, how would they design fonts for the new equipment?" I didn't expect to be capable of finding the exact answer to this question, of course, but I did feel that it would lead me in the right direction, so I began to read about the history of type design.

Well, this is a most fascinating subject, but I can't talk much about it in a limited time. Two of the first things I read were autobiographical notes by two well-known 20th century type designers, Hermann Zapf [**51**] and Frederic W. Goudy [**20**], and I was especially interested by some of Zapf's remarks:

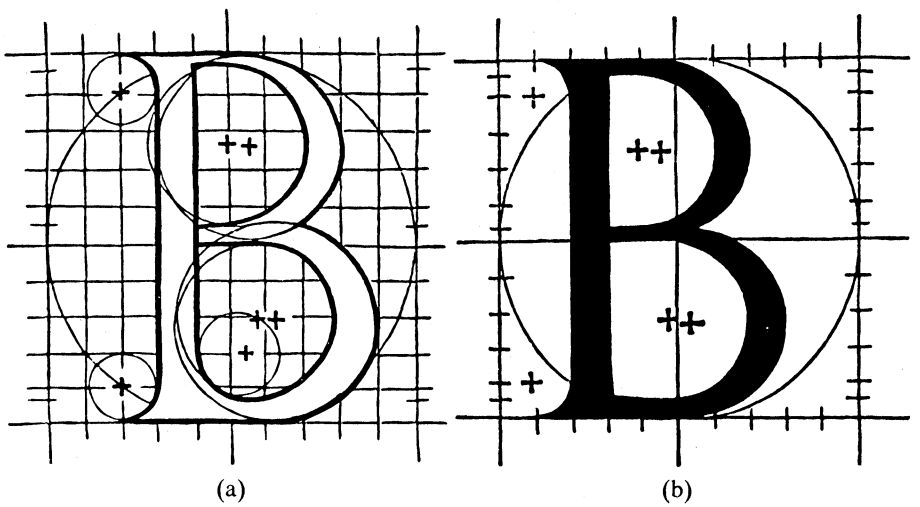
> With the beginning of the 'sixties ... I was stimulated by this new field [photocomposing] ... The type-designer—or better, let us start calling him the alphabet designer—will have to see his task and his responsibility more than before in the coordination of the tradition in the development of letterforms with the practical purpose and the needs of the advanced equipment of today .... The new photocomposing systems using cathode-ray tubes (CRT) or digital storage for the alphabet bring with them some absolutely new technical problems, many more than did the past ... [**51**, p. 71].

I have the impression that Goudy would not have been so sympathetic to the new-fangled equipment, yet his book also gave helpful ideas.



(a) (b) (c)

FIGURE 4. Sixteenth century ruler-and-compass constructions for the letter B by (a) Pacioli [42], (b) Torniello [48], and (c) Palatino [43].



(a) (b)

FIGURE 5. Two more B's, by Tory [49].

**Mathematical type design.** Fortunately the Stanford Library has a wonderful collection of books about printing, and I had the chance to read many rather rare source materials. I learned to my surprise that the idea of defining letters mathematically is by no means new, it goes back to the fifteenth century and it became rather highly developed in the early part of the sixteenth. This was the time when there were Renaissance men who combined mathematics with the real world, and in particular there was an interest in

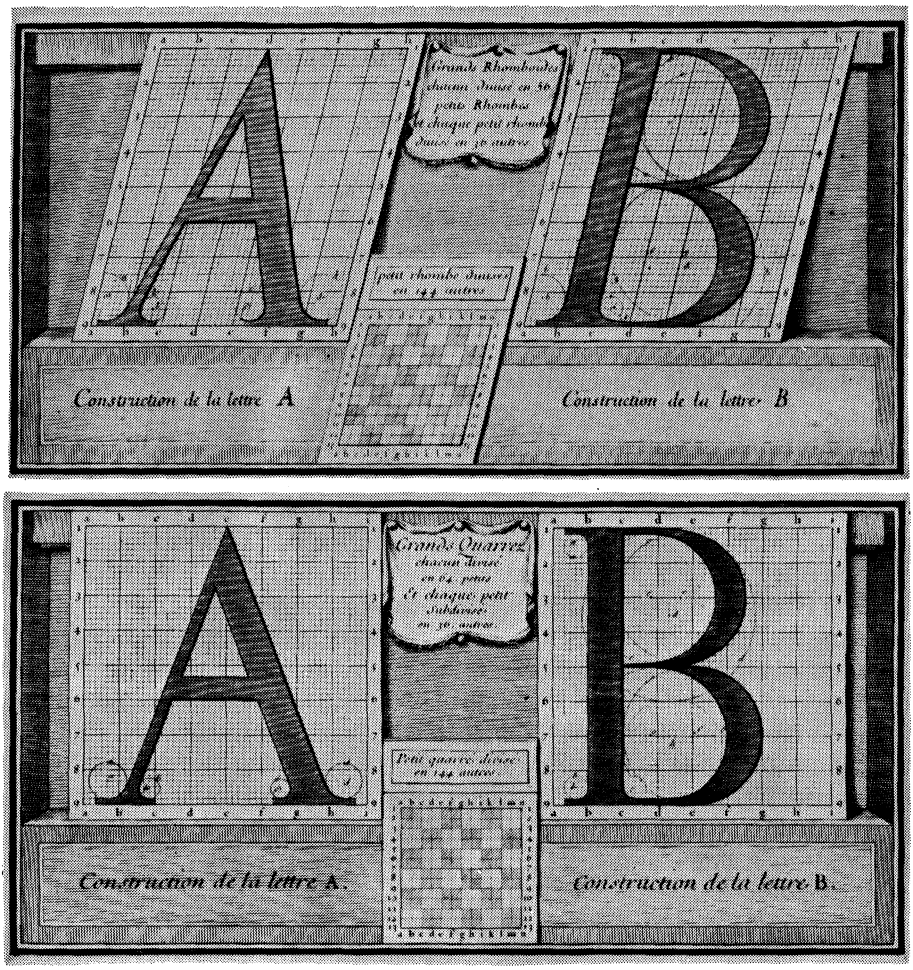



FIGURE 6. Roman and italic letters designed for Louis XIV of France [**24**].

constructing capital letters with ruler and compass. The first person to do this was apparently Felice Feliciano, about 1460, whose handwritten manuscript in the Vatican Library was published 500 years later [**19**]. Feliciano was an excellent calligrapher who wanted to put the principles of calligraphy on a sound mathematical foundation. Several other fifteen-century authors made similar experiments ([**8**] gives a critical summary of these early developments), but the most notable work of this kind appeared in the early sixteenth century.

The Italian mathematician Luca Pacioli, who had previously written the most influential book on algebra at the time (one of the first algebra books ever published), included an appendix on alphabets in his *De Divina Proportione*, a book about geometry and the "golden section" which appeared in 1509. Another notable Italian work on the subject was published by Francesco Torniello in 1517 [**48**], [**33**]; Figure 4 illustrates the letter B as constructed by Pacioli, Torniello, and by Giovanbattista Palatino [**43**]. Palatino was one of the best calligraphers of the century, and he did this work about

1550. Similar work appeared in Germany and France; the German book was probably the most famous and influential, it was Albrecht Dürer's *Underweysung der Messung* [**18**], a manual of instruction in geometry for Renaissance painters. The French book was also rather popular, it was *Champ Fleury* by Geofroy Tory [**49**], the first royal printer of France and the man who introduced accented letters into French typography. Figure 5 shows Tory's two suggestions for the letter *B*. Of all these books I much prefer Torniello's, since he was the only one who stated the constructions clearly and unambiguously.

Apparently nobody carried this work further to lower case letters, numerals, or italic letters and other symbols, until more than 100 years later when Joseph Moxon made a detailed study of some beautiful letters designed in Holland [**38**]. The ultimate in refinement of this mathematical approach took place shortly afterwards when Louis XIV of France commisioned the creation of a Royal Alphabet. A commision of artists and typographers worked on Louis's project for more than ten years beginning about 1690, and they made elaborate constructions such as those shown in Figure 6 [**24**].

Thus it is clear that the mathematical definition of letter forms has a long history. However, I must also report near-universal agreement among today's scholars of typography that those efforts were a failure. At worst, the ruler-and-compass letters have been called "ugly" and at best they are said to be "deprived of calligraphic grace" [**8**]. The French designs were not really followed faithfully by Phillipe Grandjean who actually cut Louis XIV's type, nor by anybody else to date, and F. W. Goudy's reaction to this was: "God be praised!" [**20**, p. 139]. Such strictly geometric letter forms were in fact criticized already in the sixteenth century by Giovan Cresci, a noted scribe at the Vatican Library and the Sistine Chapel. Here is what Cresci wrote in 1560:

> I have come to the conclusion that if Euclid, the prince of geometry, returned to this world of ours, he would never find that the curves of the letters could be constructed by means of circles made with compasses. [**16**]

Well, Cresci was right. But fortunately there have been a few advances in mathematics during the last 400 years, and we now have some other tricks up our sleeves besides straight lines and circles. In fact, it is now possible to prescribe formulas that match the nuances of the best type designers; and perhaps a talented designer working with appropriate mathematical tools will be able to produce something even better than we now have.

**Defining good curves.** Let's consider the following mathematical problem: Given $n$ points $z_1, z_2, \ldots, z_n$ in the plane, what is the most pleasing closed curve that goes through them in the specified order $z_1, z_2, \ldots, z_n$ and then returns to $z_1$? To avoid degenerate situations we may assume that $n$ is at least 4. This problem is essentially like the dot-to-dot puzzles that we give to young children.

Of course it is not a well-posed mathematical problem, since I didn't say what it means for a curve to be "most pleasing". Let's first postulate some axioms that the most pleasing curve should satisfy.

PROPERTY 1 (INVARIANCE). If the given points are rotated, translated, or expanded, the most pleasing curve will be rotated, translated, or expanded in the same way. [In symbols: $MPC(az_1 + b, \ldots, az_n + b) = aMPC(z_1, \ldots, z_n) + b$.]

PROPERTY 2 (SYMMETRY). Cyclic permutation of the given points does not change the solution. [$MPC(z_1, z_2, \ldots, z_n) = MPC(z_2, \ldots, z_n, z_1)$.]

PROPERTY 3 (EXTENSIONALITY). Adding a new point that is already on the most pleasing curve does not change the solution. [If $z$ is between $z_k$ and $z_{k+1}$ on $MPC(z_1, \ldots, z_n)$ then $MPC(z_1, \ldots, z_k, z, z_{k+1}, \ldots, z_n) = MPC(z_1, \ldots, z_k, z_{k+1}, \ldots, z_n)$.]

These properties are rather easy to justify on intuitive grounds. For example, the extensionality property says that additional information won't lead to a poorer solution.

The next property is not so immediately apparent, but I believe it is important for the application I have in mind.

PROPERTY 4 (LOCALITY). Each segment of the most pleasing curve between two of the given points depends only on those points and the ones immediately preceding and following. [$MPC(z_1, z_2, \ldots, z_n)$ is composed of $MPC(z_n, z_1, z_2, z_3)$ from $z_1$ to $z_2$, then $MPC(z_1, z_2, z_3, z_4)$ from $z_2$ to $z_3, \ldots$, then $MPC(z_{n-1}, z_n, z_1, z_2)$ from $z_n$ to $z_1$.]

According to the locality property, changes to one part of a pattern won't affect the other parts. This simplifies our search for the most pleasing curve, because we need only solve the problem in the case of four given points; and experience shows that it is also a great simplification when letters are being designed, since individual portions of strokes can be dealt with separately. Incidentally, Property 4 implies Property 2 (cyclic symmetry).

One way to satisfy all four of these properties is simply to let the most pleasing curve consist of straight line segments. But this doesn't seem adequately pleasing, so we postulate

PROPERTY 5 (SMOOTHNESS). There are no sharp corners in the most pleasing curve. [$MPC(z_1, \ldots, z_n)$ is differentiable, under some parameterization.]

In other words, there is a unique tangent at every point of the curve.

The extensionality, locality, and smoothness properties taken together imply, in fact, that *the direction of the tangent at $z_k$ depends only on $z_{k-1}$, $z_k$ and $z_{k+1}$*. For this tangent appears in two curves, the one from $z_{k-1}$ to $z_k$ and the one from $z_k$ to $z_{k+1}$, hence we know that it depends only on $(z_{k-2}, z_{k-1}, z_k, z_{k+1})$ and that it depends only on $(z_{k-1}, z_k, z_{k+1}, z_{k+2})$. By the extensionality property, we can assume that $n$ is at least 5, so $z_{k-2}$ is different from $z_{k+2}$ and the tangent must be independent of them both. We have actually used only a very weak form of extensionality in this argument.

If we apply the full strength of the extensionality postulate, we obtain a much stronger consequence, which is quite unfortunate: *There is no good way to satisfy Properties* 1–5! For example, suppose we add one more axiom, which is almost necessary in any reasonable definition of pleasing curves:

PROPERTY 6 (ROUNDNESS). If $z_1$, $z_2$, $z_3$, $z_4$ are consecutive points of a circle, the most pleasing curve through them is that circle.

This property together with our previous observation about the tangent depending only on three points completely determines the tangent at each of our given points; namely, the tangent at $z_k$ is the tangent to the circle which passes through $z_{k-1}$, $z_k$, and $z_{k+1}$. (Let's ignore for the moment the possibility that these three points lie on a straight line.) Now the extensionality property says that if $z$ is any point between $z_1$ and $z_2$ on the most pleasing curve for $z_1, \ldots, z_n$, we know the tangent direction at $z$, as long as $z$ is not on the line from $z_1$ to $z_2$. But there is a unique curve starting at any $z$ off this line and having the specified tangents at each of its points, namely the arc of the circle from $z$ to $z_2$ passing through $z_1$: No matter where you start, off the straight line, there is only one curve having the correct tangents. It follows that the tangent at $z_2$ depends only on $z_1$, $z_2$, and the tangent at $z_1$, and this is impossible.

The above argument proves that there is no way to satisfy Properties 3, 4, 5, and 6. A similar argument would show the impossibility for any reasonable replacement for Property 6, since the tangents determined for all $z$ between $z_1$ and $z_2$ will define a vector field in which there are unique curves through essentially all of the points $z$, yet a two-parameter family of curves is required between $z_1$ and $z_2$ in order to allow sufficient flexibility in the derivatives there.

So we have to give up one of these properties. The locality property is the most suspicious one, but I mentioned before that I didn't want to give it up; therefore the extensionality property has to go. This means that if we take the most pleasing curve through $z_1, \ldots, z_n$ and if we specify a further point $z$ actually on this curve between $z_{k-1}$ and $z_k$, where the tangent at $z$ is not the same as the tangent to the circle from $z_{k-1}$ to $z$ to $z_k$, then the "most pleasing" curve through these $n + 1$ points will be different. A possible virtue is that we are encouraged not to specify too many points; a possible drawback is that we may not be able to get the curves we want.

**A practical approximation.** Returning to the question of type design, our goal is to specify a few points $z_k$ and to have a mathematical formula that defines a pleasant curve through these points; such curves will be used to define the shape of the character we are designing. Ideally it should also be easy to compute the curves. I decided to use cubic equations

$$z(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$$

where $\alpha_0$, $\alpha_1$, $\alpha_2$, $\alpha_3$ are complex numbers and $t$ is a real parameter. The curves I am dealing with are *cubic splines*, namely piecewise cubic equations, since a different cubic will be used in each interval between two of the given points; however, the way I am determining the coefficients of these two cubics is different from any of the methods known to me, in my limited experience with the vast literature about splines. Perhaps my way to choose the coefficients is more awkward than the usual ones; but I have obtained good results with it, so I'm not ashamed to reveal the curious way I proceeded.

In the first place, I decided that the cubic equation between $z_1$ and $z_2$ should be determined completely by $z_1$ and $z_2$ and the directions of the tangents at $z_1$ and $z_2$. We have already seen that these tangents are essentially predetermined if Properties 4, 5, and 6 are to be valid, and I have also found frequent occasion in type design when it was desirable to specify that a certain tangent was to be made horizontal or vertical. Thus, my method of computing a nice curve through a given sequence of points is first to compute the tangent directions at each point, then to compute the cubics in each interval based solely on the endpoints of that interval and on the desired tangents there. By rotation and translation and scaling, according to Property 1, we can assume that the problem is to go in the complex plane from 0 to 1, with given directions at the endpoints. The most general cubic equation which does this is

$$z(t) = 3t^2 - 2t^3 + re^{i\theta}t(1-t)^2 - se^{-i\varphi}t^2(1-t),$$

and it remains to determine positive numbers $r$ and $s$ as appropriate functions of $\theta$ and $\varphi$.

In the second place, I realized that it was impossible to satisfy Property 6 with cubic splines, because you can't draw a circle as a cubic function of $t$. But I wanted to be able to get curves that were as near to being circles as possible, whenever four consecutive data points lay on a circle; the curves should preferably be indistinguishable from circles as far as the human eye is concerned. Therefore when $\theta = \varphi$ I decided to choose $r = s$ in such a way that $z(\frac{1}{2})$ was precisely on the relevant circle, hoping that the curve between 0 and $\frac{1}{2}$ and between $\frac{1}{2}$ and 1 wouldn't veer too far away. Well, this turned out to work extremely well: A little calculation, done with the help of a computer,[4] showed that the maximum deviation from a true circle occurs at the point $t = (3 \pm \sqrt{3})/6$, and the relative error is negligibly small. For example, if we take four points equally spaced at distance 1 from some center, the spline curve defined by these points in the above manner stays between distance 1 and distance $71/54 - 2\sqrt{2/9} < 1.00055$ from the center, an error of less than one part in a thousand. If there are 8 points, the maximum error is less than four parts per million; and if there are $n$ points, the maximum error goes to zero as $1/n^6$.

(Changing the notation slightly, let

$$z(t) = 1 + (e^{i\theta} - 1)(3t^2 - 2t^3) + 4it(1-t)(1-t-e^{i\theta}t)\left(\sin\frac{\theta}{2}\right)/\left(1+\cos\frac{\theta}{2}\right)$$

and $f(t) = |z(t)|^2$. Then

$$f'(t) = 8\left(\sin^2\frac{\theta}{2}\right)\left[\frac{\cos\frac{\theta}{2}-1}{\cos\frac{\theta}{2}+1}\right]^2 (t-1)t(2t-1)(6t^2-6t+1)$$

[4] Thanks are due to the developers of the computer algebra system called MACSYMA at MIT, and to the ARPA network which makes this system available for research work.

and

$$\max_{0\leqslant t\leqslant 1}|z(t)| = \left|z\left(\frac{3-\sqrt{3}}{6}\right)\right| = 1 + \frac{\theta^6}{55296} + \frac{\theta^{10}}{106168320} + \cdots,$$

while $\min_{0\leqslant t\leqslant 1}|z(t)| = z(0) = z(\frac{1}{2}) = z(1) = 1$. The "two-point circle" has $\max|z(t)| = \sqrt{28/27} = 1.01835$, while the three-point circle has $\max|z(t)| = \sqrt{325/324} = 1.001542$, and the eight-point cicle has $\max|z(t)| = 1.0000042455$.)

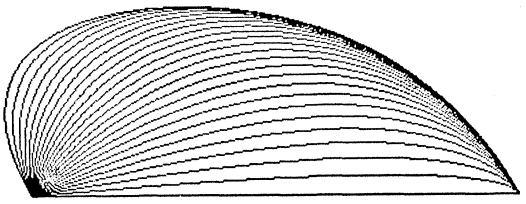


FIGURE 7.
Spline curves with $\theta = 0°$ (5°) 120° and $\phi = 60°$.

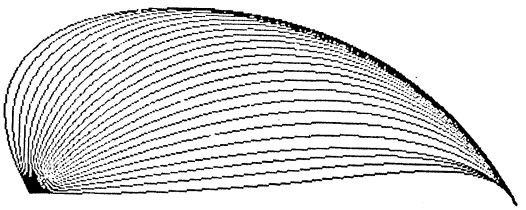


FIGURE 8.
Same as Figure 7 but adjusted so that $r' = \max(\frac{1}{2}, r)$, $s' = \max(\frac{1}{2}, s)$.

Another case when a natural way to choose $r$ and $s$ suggests itself is when $\theta + \varphi = 90°$; then the curve $z(t)$ should be nearly the same as an ellipse having the endpoints on its axes. (This boils down to requiring that $(3t^2 - 2t^3 - (s/\cos\varphi)t^2(1-t) - 1)^2 + (3t^2 - 2t^3 + (r/\cos\theta)t(1-t)^2)^2$ be approximately equal to 1.) So far therefore I knew that I wanted

$$r = \frac{4}{1+\cos\theta}, \qquad s = \frac{4}{1+\cos\varphi} \qquad \text{when } \theta = \varphi;$$

$$r = \frac{4\cos\theta}{(1+\cos 45°)(\cos 45°)}, \qquad s = \frac{4\cos\varphi}{(1+\cos 45°)(\cos 45°)}$$

when $\theta + \varphi = 90°$.

So I tried the formulas

$$r = \frac{4\cos\theta}{\left(1+\cos\frac{\theta+\varphi}{2}\right)\left(\cos\frac{\theta+\varphi}{2}\right)}, \qquad s = \frac{4\cos\varphi}{\left(1+\cos\frac{\theta+\varphi}{2}\right)\left(\cos\frac{\theta+\varphi}{2}\right)}$$

which fit both cases. However, this didn't give satisfactory results, especially when $\theta + \varphi$ approached 180°. My second attempt was

$$r = \left|\frac{4\sin\varphi}{\left(1+\cos\frac{\theta+\varphi}{2}\right)\sin\frac{\theta+\varphi}{2}}\right|, \qquad s = \left|\frac{4\sin\theta}{\left(1+\cos\frac{\theta+\varphi}{2}\right)\sin\frac{\theta+\varphi}{2}}\right|$$

and this has worked very well. Figure 7 shows the spline curves that result from the above approach when $\varphi = 60°$ and when $\theta$ varies from 0° to 120° in 5° steps.

It can be proved that if $\theta$ and $\varphi$ are nonnegative and less than 180°, the cubic curve $z(t)$ I have defined will never cross the straight lines at angles $\theta$ and $\varphi$ that meet the endpoints 0 and 1 respectively. This is a valuable property in type design, since it can be used to guarantee that the curve won't get out of bounds. However, I found that it also led to unsatisfactory curves when one of $\theta$ or $\varphi$ was very small and the other was not, since this meant that the curve $z(t)$ would be very close to a straight line yet it would enter that line from outside at a rather sharp angle. In fact, the angle $\theta$ is not infrequently zero, and this forces a straight line and a sharp corner at the right endpoint. Therefore I changed the formulas by making sure that both $r$ and $s$ are always $\frac{1}{2}$ or greater unless special exceptions are made; furthermore I never let $r$ or $s$ exceed 4. Figure 8 shows the spline curves obtained under the same conditions as Figure 7, but with $s$ set to $\frac{1}{2}$ if the above formula calls for any smaller value.

Using these techniques we obtain a system for drawing reasonably nice curves, if not the most pleasing ones, and it is especially good at circles. If the method gives the wrong tangent direction at some point, you can control this by specifying two points very close together having the desired slope. I have also included another way to modify the standard tangent directions, intended to make the system as good at drawing ellipses as it is at drawing circles: Before computing the splines I first shrink the entire figure in the vertical direction by multiplying all the $y$ coordinates by a given aspect ratio (normally 1); then the splines are calculated, and the resulting shrunken curves are stretched out again by dividing the $y$ coordinates by the aspect ratio.

**Application to type design.** Now let's take a closer look at what can be drawn with a mathematical system like this. I suppose the natural thing to show you would be the letters $A$ to $Z$; but since this is a mathematical talk, let's consider the digits 0 to 9 instead. (See Figure 9.) Incidentally, the way I have arranged these numerals illustrates a fundamental distinction between a mathematician and a printer: the mathematician puts 0 next to the 1, but the printer always puts it next to the 9.


FIGURE 9. Digits 0 to 9 drawn by the prototype METAFONT programs.
(Further refinements to these characters will be made before the font has its final form.)

Most of these digits are drawn by using another idea taken from the history of typography, namely to imitate the calligrapher who uses pen and ink. Consider first the numeral '3', for example. The computer program which drew this symbol in Figure 9 can be paraphrased as follows. "First draw a dot whose left boundary is $\frac{1}{6}$ of the way from the left edge to the right edge of the type and whose bottom boundary is $\frac{3}{4}$ of the way from the top to the bottom of the desired final shape. Then take a hairline pen and, starting at the left of the dot, draw the upward arc of an ellipse; after reaching the top, the pen begins to grow in width, and it proceeds downward in another ellipse in such a way that the maximum width occurs on the axis of the ellipse, with the right

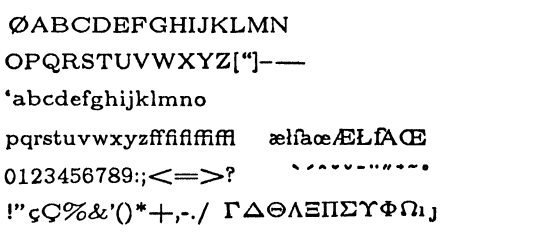



FIGURE 10. A font of 128 characters defined by METAFONT with standard pen settings. (The accent characters will be appropriately raised and centered over other letters when used by TEX.)

edge of the pen $\frac{8}{9}$ of the way from the left edge to the right edge of the type. Then the pen width begins to decrease to its original size again as the pen traverses another ellipse taking it down to a position 48% of the way from the top to the bottom of the desired final shape. . . . ”

Notice that instead of describing the boundary of the character, as the renaissance geometers did, my METAFONT system describes the curve traveled by the *center* of the *pen*, and the shape of this pen is allowed to vary as the pen moves. The main advantage of this approach is that the same definition readily yields a family of infinitely many related fonts of type, each font being internally consistent. The change in pen size is governed by cubic splines in a manner analogous to the motion of the pen's center. In order to define the 20 or so different type fonts used in various places in my books, I need for the most part to use only three kinds of pens, namely (i) a circular pen, used for example to draw dots and at the base of the numeral ‘7’; (ii) a horizontal pen, whose shape is an ellipse, the width being variable but the height being constantly equal to the height of a hairline pen—such a pen is used most of the time, and in particular to draw all of the numeral ‘3’ except for the dots; (iii) a vertical pen, analogous to the horizontal one, used for example to draw the strokes at the bottom of the ‘2’ and at the top of the ‘5’ and the ‘7’. For the fonts I am using, it was not necessary to use an oblique pen (i.e., an ellipse that is tilted on its side) except to make the tilde accent for Spanish *n*’s; but to produce fonts of type analogous to Times Roman, an oblique pen would of course be used. If this system were to be extended to Chinese and Japanese characters, I believe it would be best to add another degree of freedom to the pen's motion, allowing an elliptical pen shape to rotate as well as to change its width.

The digit ‘4’ shows another aspect of the METAFONT system. Although this character is fairly simple, consisting entirely of straight lines, notice that the thick line has to be cut off at an angle at the top. In order to do this, there are *erasers* as well as pens. First the computer draws a thick line all the way from top to bottom, like the upper case letter ‘*I*’, then it takes an eraser which erases everything to its left and comes down the diagonal stroke, then it takes a hairline pen and finishes the diagonal stroke. Such an eraser is used also at the top of the ‘1’ and the bottom of the ‘2’, etc.

Sometimes a simple spline seems to be inadequate to describe the proper growth of pen width, so in a few cases I had to resort to describing the left and right edges of the pen as separate curves, to be filled in afterwards. This

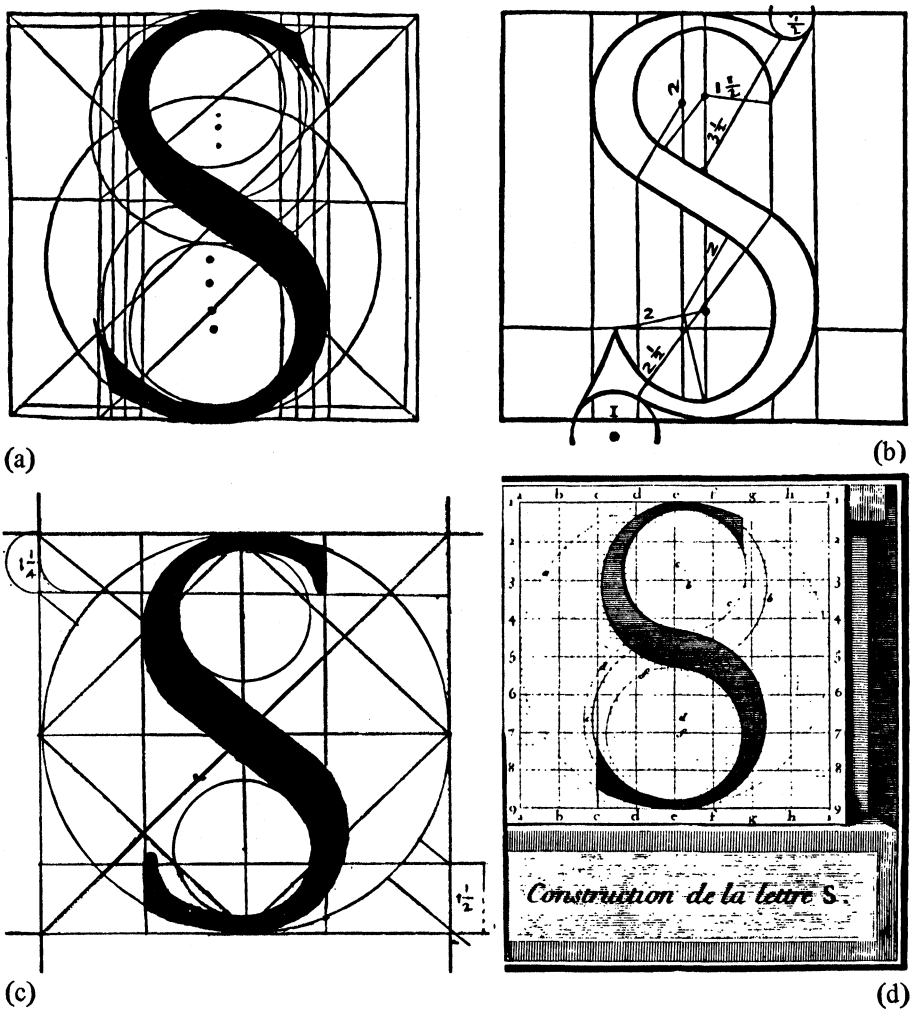



FIGURE 11. The letter *S* as defined by (a) Pacioli [**42**]; (b) Torniello [**48**]; (c) Palatino [**43**]; (d) French commission under Jaugeon [**24**].

occurs for example in the main stroke of the numeral '2', whose edges are defined by two splines having a specified tangent at the bottom and having vertical slope at the right of the curve.

With these techniques I found that it was possible to define a decent-looking complete font, containing a total of 128 characters, in about two months, although it will still be necessary of course to do fine tuning when more trial pages are typeset. (See Figure 10.) The most difficult symbol by far, at least for me, was the letter *S* (and the numeral 8, which uses the same procedure); in fact I spent three days and nights without sleep, trying to make the *S* look right, before I got it. At one point I even felt it would be easier to rewrite all my books without using any *S*'s! After the first day of discouraging trials, I showed what I had to my wife, and she said, "Why don't you make it *S*-shaped?"

Figure 11 shows how this problem was solved by Pacioli, Torniello, Palatino, and the French academicians; but the letter doesn't look like a

modern *S*. Furthermore I think the engraver of the French *S* cheated a little in rounding off some lines near the middle—perhaps he used a French curve. With my wife's assistance, I finally came up with a satisfactory solution, somewhat like those used in the sixteenth century but generalized to ellipses. Each boundary of each arc of my *S* curve is composed of an ellipse and a straight line, determined by (i) the locations of the beginning and ending points, (ii) the slope of the straight line, and (iii) the desired left extremity of the curve. It took me three hours to derive the necessary formulas, and I think Newton and Leibnitz would have enjoyed working on this problem. Figure 12 shows various trial *S*'s drawn by this scheme with different slopes; I hope you prefer the middle one, since it is the one I am actually using.

FIGURE 12. Different S's obtained by varying the slope in the middle. (This shows $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, 1, $\frac{4}{3}$, $\frac{3}{2}$, and 2 times the "correct" slope.)

**Families of fonts.** To extend the METAFONT system, one essentially writes a computer program for the description of each character, in a special language intended for describing pen and eraser strokes. My colleague R. W. Gosper has observed that this is the opposite of *Sesame Street*: Instead of "This program was brought to you by the letter *S*" we have "This letter *S* was brought to you by a program." There are about 20 parameters to the program, telling how big a hairline pen is, how wide it should be when drawing straight or curved stem lines, and specifying the sizes and proportions of various parts of the letters (the *x*-height, the heights of ascenders and descenders, the set width, the length of serifs, and so forth). By changing these parameters, we obtain infinitely many different styles of type, yet all of them are related and they seem to blend harmoniously with each other.

For example, Figure 13 shows some of the possibilities. In Figure 13a we have a conventional "modern" font in the tradition of Bodoni and Bell and "Scotch Roman". Then Figure 13b shows a corresponding boldface, in which the hairlines are slightly larger and the stem lines are substantially wider. By making the hairlines and stem lines both the same size, and setting the serif length to zero, we obtain a sans-serif font as shown in Figure 13c. All of these examples are produced with the same programs defining the letter shapes; only the parameters are being varied. Actually the particular font shown in Figure 13c will have a different style of *g*, because the descenders are especially short in this font, but I have shown this "*g*" in order to illustrate the parametric variations. Figure 13d shows a boldface sans-serif style in which the pen has an oval shape wider than it is tall; I find this style especially pleasing, particularly because it came out by accident—I designed the programs only so that two or three different fonts would look right, all the others are free bonuses, and I had no idea that this one would be so nice.

With a suitable setting of the parameters, we can even imitate a typewriter with its fixed width letters, as shown in Figure 13e. There is also a provision

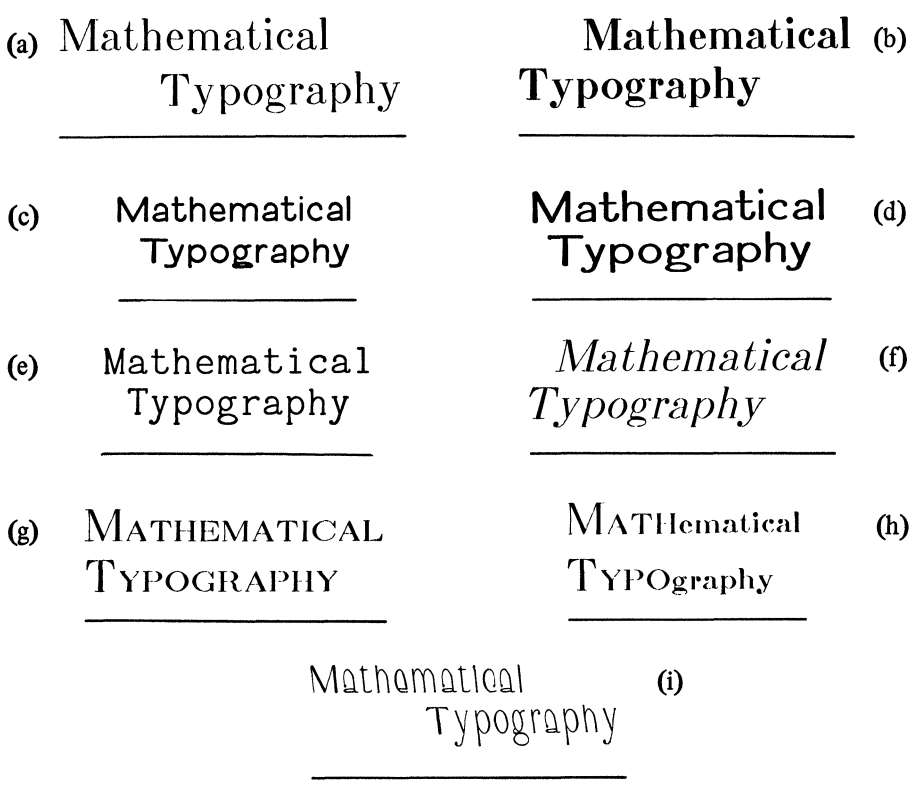



FIGURE 13. Different styles of type obtained by varying the parameters to METAFONT:
(a) Computer modern roman; (b) Computer modern bold;
(c) Computer modern san serif; (d) Computer modern sans serif bold;
(e) Computer modern typewriter; (f) Computer modern slanted roman;
(g) Computer modern roman with small caps; (h) Computer modern roman with small caps and "small lower case";
(i) Computer modern funny.

to slant the letters as in Figure 13f; here the pen position is varied, but the actual shape of the pen is not being slanted, so circles remain circles.

Another setting of the parameters leads to caps and small caps as shown in Figure 13g; small caps are drawn with the pens and heights ordinarily used for lower case letters, but controlled by the programs for upper case letters. Figure 13h shows something printers have never seen before: this is what happens when you draw lower case letters in the small caps style, and we might call it "small lower case". It actually turns out to be one of the most pleasing fonts of all, except that the dots are too large.

Finally, Figure 13i illustrates the variations you can get by giving weirder settings to the parameters.

When I was an assistant professor at Caltech, the math department secretaries used to send occasional "crank" visitors to my office, and I recall one time when a man came to ask if anybody had calculated the value of $\pi$

"out to the end" yet. I tried to explain to him that $\pi$ had been proved irrational, but this didn't seem to sink in, so finally I showed him a table of $\pi$ to 100,000 decimals and told him that the expansion hadn't ended yet. I wish I could have had my typographical system ready at that time, so that I could have shown him Figure 14!

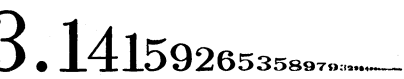



FIGURE 14. Variation in height, width, and pen size.

Figure 14 illustrates another principle of type design, namely that different sizes of type in the same style are not simply obtained from each other by optical transformations. The heights and widths and pen stroke sizes change at different rates, and a good typographer will design each size of type individually. I'm not claiming that Figure 14 shows the best way for the proportions to vary; it will take further experimentation before I have a good idea of what is desirable. The point I wish to make is that the alteration of type sizes for subscripts and so on is not as simple as it might seem at first, but a system like METAFONT will be able to vary the parameters quite readily, and visual experiments on different parameter settings can be carried out quickly. It used to take months for a type designer to make his drawings and have them converted to metal molds before he could see any proofs. One of the results was that there simply wasn't time to give proper attention to all the mathematical symbols and Greek letters, etc., as well as to the more common symbols, so a printer of mathematics had to make do with a hodge-podge of available characters in different sizes. (For example, he was often obliged to use different styles of letters in subscript positions, as we have seen.) Under the approach I am recommending, we automatically get consistency of all the symbols whenever the parameters change.

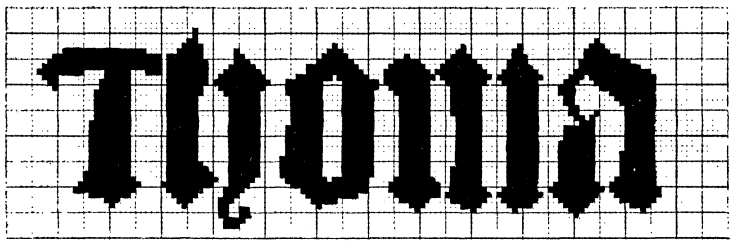


FIGURE 15. Lettering equivalent to this raster pattern appears in a Norwegian tapestry from Gildeskaal old church, woven about 1500 [22, p. 116].

**From continuous to discrete.** The METAFONT system must not only define the characters in the continuum on the plane, it must also express them in terms of a discrete raster. Such squaring off of letters on graph paper has a long history, going back far before the invention of computers or television; for example, we all can remember seeing cross-stitch embroidery samplers from the nineteenth century. The same idea on a finer scale has been used in tapestries for many centuries: In our own home library, my wife found the example of Figure 15 which was woven in the northern part of Norway about

1500; this shows the name of St. Thomas in a style imitating contemporary calligraphy, and I'm sure that examples which antedate the printing press can be found elsewhere.

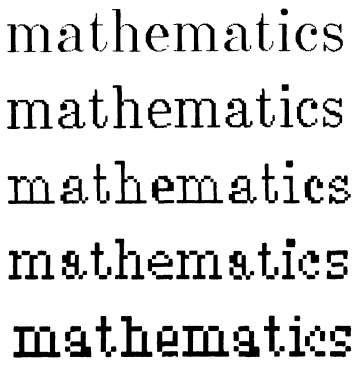



FIGURE 16.
Adjusting the letters to coarser rasters.

Figure 16 shows how METAFONT produces the same letters from the same parameters but with different degrees of resolution in the raster. This digitization process itself is considerably more difficult than it may seem at first, and some nontrivial mathematical concepts were needed before I could produce satisfactory results. In the first place, it is not sufficient merely to draw or to imagine drawing the character with infinite precision and then to "round" it by blacking in all the squares on graph paper that are sufficiently dark in the true image. One of the reasons this fails is that the three stem lines of the *m*, for instance, might be located in different relative positions with respect to the grid, so that the first stroke might round to three units wide (say) and the second might round to four. This would be quite unsatisfactory, as the eye quickly picks up such a variation in thickness, but it is avoided by METAFONT since the pen itself is first digitized and then the same digitized pen is used for all three strokes. Another problem is that those three strokes should be equally spaced; it would look bad if there were seven units between the first two and eight units between the last two, so the program for '*m*' needs to round its points in such a way that this doesn't happen.

The process of digitizing the pen is not trivial either. Suppose, for example, we want a circular pen that is 2 raster units wide; the appropriate pen is clearly a $2 \times 2$ square, which is the closest to a circle that we can come at this low degree of resolution. Now notice that we can't *center* a $2 \times 2$ square on any particular square, since none of the four squares is at its center; the same problem arises whenever we have to deal with a pen having even dimensions. One way to resolve this would be to insist on working only with odd numbers, but this would be far too limiting; so METAFONT uses a special rounding rule for the position of the pen's center. In general, suppose the pen is an ellipse of integer width $w$ and integer height $h$; then if the pen is to be positioned at the real coordinates $(x, y)$, its actual position on the discrete grid is taken to be

$$(\lfloor x - \delta(w) \rfloor, \lfloor y - \delta(h) \rfloor)$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x$, and $\delta(\text{even}) = \frac{1}{2}$, $\delta(\text{odd}) = 0$. The pen itself, if positioned at the origin, would

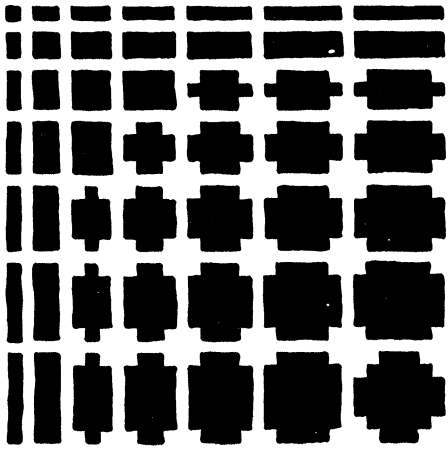
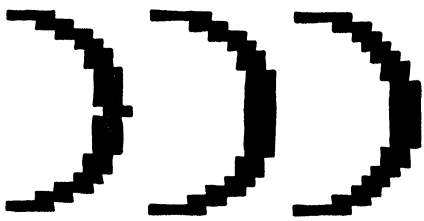


FIGURE 17.
Discrete "elliptical" pens
of small integer width and height.

consist of all integers $(x, y)$ which satisfy

$$\left(\frac{2(x-\delta(w))}{2}\right)^2+\left(\frac{2(y-\delta(h))}{2}\right)^2 \leqslant 1+\max\left(\frac{2\delta(w)}{w}, \frac{2\delta(h)}{h}\right)^2.$$

This formula—which incidentally is not the first one I tried—ensures that the discrete pen will indeed be $w$ units wide and $h$ units high, when $w$ and $h$ are positive integers. Figure 17 shows the pens obtained for small $w$ and $h$.



(a) (b) (c)

FIGURE 18.
Difficulties of rounding an arc
properly. (Three circles of radius
10 drawn with a $1 \times 3$ pen.)

Still another problem appears when we want curved lines to look right. Figure 18(a) shows a semicircle of radius 10 units, drawn with a pen of height 1 and width 3, when the right boundary of the pen falls exactly at an integer point; the pen sticks out terribly in one place. On the other hand if this right boundary falls just shy of an integer point, we get the curve in Figure 18(b) which looks too flat. The ideal occurs in Figure 18(c), when the right boundary occurs exactly midway between integers. Therefore the METAFONT programs adjust the location of curves to the raster before actually drawing the curves, forcing the favorable situation of Figure 18(c); the actual shape of each letter changes slightly in order to adapt that letter to the desired raster size in a pleasant way.

There is yet another problem, which arises when the pen is growing in such a way that the edges of the curve it traces would be monotonic if the pen were drawn to infinite precision, yet the independent rounding of pen location and pen width causes this monotonicity to disappear. The problem arises only rarely, but when it does happen the eye immediately notices it.

| Pen width and location | Rounded width and location |
|---|---|
| (3.5, 0.5,10.5) | (3, 0,10) |
| (3.7, 0.8, 9.5) | (3, 0, 9) |
| (3.9, 1.1, 8.5) | (3, 1, 8) |
| (4.1, 1.4, 7.5) | (4, 0, 7) |
| (4.3, 1.7, 6.5) | (4, 1, 6) |
| (4.5, 2.0, 5.5) | (4, 1, 5) |
| (4.7, 2.3, 4.5) | (4, 1, 4) |
| (4.9, 2.6, 3.5) | (4, 2, 3) |
| (5.1, 2.9, 2.5) | (5, 2, 2) |
| (5.3, 3.2, 1.5) | (5, 3, 1) |
| (5.5, 3.5, 0.5) | (5, 3, 0) |

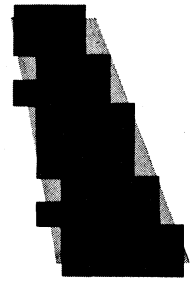


FIGURE 19. Failure of monotonicity due to independent rounding. (Rounding takes $(w, x, y)$ into $(\lfloor w \rfloor, \lfloor x - \delta(\lfloor w \rfloor) \rfloor, \lfloor y \rfloor)$.)

Consider, for example, the completely linear situation in Figure 19, where each decrease by one unit in $y$ is accompanied by an increase of .3 units in $x$ and an increase of .2 units in the pen width $w$; the intended pen height is constant and very small, but in the discrete case the pen height is taken to be 1. The lightly shaded portion of Figure 19 shows the true shape intended, but the darker squares show that the digitized form yields a nonmonotonic left boundary. METAFONT compensates for this sort of problem by keeping track of the desired boundaries when the pen width is varying, plotting points twice (e.g. plotting both $(x, y)$ and $(x - 1, y)$) when necessary to keep the boundary correct. In other words, the idea of rounding the pen location and the width independently is sometimes effectively abandoned.

The final digitization problem that I needed to resolve was to make the left half of an "0" look like the mirror image of its right half, to make a left parenthesis look like the mirror image of a right parenthesis, and so on. This was done by having the METAFONT programs in such cases choose a center point that was either exactly at an integer or an integer plus $\frac{1}{2}$, and to introduce dual rounding which could be proved to produce exactly the correct symmetry properties.

**Alternative approaches.** As I have said, I believe the METAFONT system is successful as a way to define letters and other symbols, but probably even better procedures can be divised with further research. Some of the limitations of my cubic splines are indicated in Figure 20. Part (a) of that illustration shows a five-pointed star and the word "mathematics" in an approximation to my own handwriting, done with straight line segments so that you can see exactly what the data points are that I fed to my spline routine. Part (b) shows the way my handwriting might look when I get older; it was obtained by simply setting $r = s = 2$ in all the spline segments, therefore making clear what tangent angles are prescribed by the system. Part (c) is somewhat more disciplined, it was obtained by putting $r = s = \frac{1}{2}$ everywhere. Figure 20(d) is like Figure 20(c) but drawn with a combined pen-and-eraser. Such a combination can lead to interesting effects, and the star here is my belated contribution to America's bicentennial.

When the general formulas for cubic splines are used as I explained above, we get Figure 20(e) in which the star has become a very good approximation

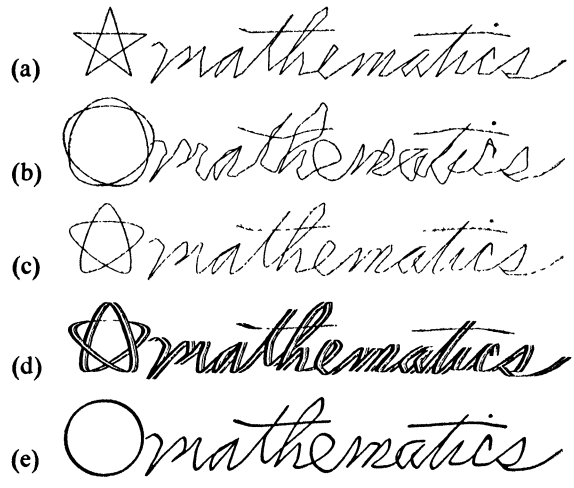



FIGURE 20.
Examples of the cubic splines applied to sloppy handwriting.

to a circle (as I said it would). In this illustration the pen is thicker and has a slightly oblique stress. Although my handwriting is inherently unbeautiful, there are still some kinks in Figure 20(e) that could probably be ironed out if a different approach were taken.

The most interesting alternative from a mathematical standpoint seems to be to find a curve of given length that minimizes *the integral of the square of the curvature* with respect to arc length. This integral is proportional to the strain energy in a mechanical spline (in other words, a thin slat or beam) of the given length, going through the given points, so it seems to be an appropriate quantity to minimize. E. H. Lee and G. E. Forsythe [31] have reviewed early work on this variational problem, and shown that it is equivalent to having the spline at equilibrium with forces applied only at the given points of support. The Norwegian mathematician Even Mehlum [36] has shown that if we specify a fixed arc length between consecutive points, the optimum curve will have linearly changing curvature of the form $ax + by + c$ at point $(x, y)$, and he has suggested choosing the constants by taking $b/a = (y_2 - y_1)/(x_2 - x_1)$ between $(x_1, y_1)$ and $(x_2, y_2)$, and requiring that slope and curvature be continuous across endpoints. Such an approach seems to require considerably more computation than the cubic splines recommended here, but it may lead to better curves, e.g. satisfying the extensionality property.

Another interesting approach to curve-drawing, which may be especially useful for simulating handwriting, is a "filtering" method suggested to me recently by Michael S. Paterson of the University of Warwick (unpublished). To get a smooth curve passing through points $z_k$, assuming that these points are about equally spaced on the desired curve, one simply writes

$$z(t) = \sum_k (-1)^k z_k f(t - k) \Big/ \sum_k (-1)^k f(t - k)$$

where $f(t)$ is an odd function of order $t^{-1}$ as $t \to 0$, decreasing rapidly away from zero; e.g.,

$$f(t) = \operatorname{csch} t = 2/(e^t - e^{-t}).$$

I have not had time yet to experiment with Paterson's method or to attempt to harness it for the drawing of letters. It is easy to see that the derivative $z'(z_k) = f(1)(z_{k+1} - z_{k-1}) - f(2)(z_{k+2} - z_{k-2}) + \cdots$ lies approximately in the direction of $z_{k+1} - z_{k-1}$.

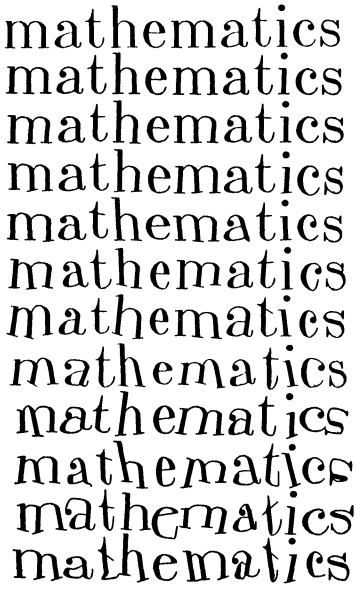



FIGURE 21.
Increasingly random pen positions; $\sigma = 0, 1, \ldots$ .

**Randomization.** In conclusion, I'd like to report on a little experiment that I did with random numbers. One might complain that the letters I have designed are too perfect, too much like a computer, so they lack "character." In order to counteract this, we can build a certain amount of randomness into the choices of where to put the pen when drawing each letter, and Figure 21 shows what happens. The coordinates of key pen positions were chosen independently with a normal distribution and with increasing standard deviation, so that the third example has twice as much standard deviation as the second, the fourth has three times as much, and so on. Note that the two *m*'s on each line (except the first) are different, and so are the *a*'s and the *t*'s, since each letter is randomly drawn.

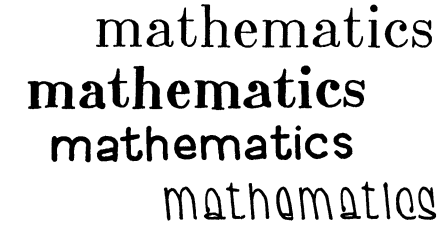



FIGURE 22.
A bit of randomness introduced into various styles of type.

After the deviation gets sufficiently large the results become somewhat ludicrous; and I don't want people to say that I ended this lecture by making a travesty of mathematics. So let us conclude by looking at Figure 22, which shows what is obtained in various fonts when the degree of randomness is somewhat controlled. I think it can be said that the letters in this final example have a warmth and charm which makes it hard to believe that they

were really generated by a computer following strict mathematical rules. Perhaps the reason that the printing of mathematics looked so nice in the good old days was that the fonts of type were imperfect and inconsistent.

**Summary.** I'd like to summarize now by pointing out the moral of this long story. My experiences during the last few months vividly illustrate the fact that there are plenty of good mathematical problems still waiting to be solved, almost everywhere you look—especially in areas of life where mathematics has rarely been applied before. Mathematicians can provide solutions to these problems, receiving a double payoff—namely the pleasure of working out the mathematics, together with the appreciation of the people who can use the solutions. So let's go forth and apply mathematics in new ways.

**Acknowledgments.** I would like to thank my wife Jill for the many important suggestions she made to me during critical stages of this research; also Leo Guibas and Lyle Ramshaw for the help they provided in making illustrations at Xerox Palo Alto Research Laboratories; also Lester Earnest, Michael Fischer, Frank Liang, Tom Lyche, Albert Meyer, Michael Paterson, Michael Plass, Bob Sproull, Jean E. Taylor, and Hans Wolf, for helpful ideas and stimulating discussions and correspondence about this topic; also Gordon L. Walker, for verifying my conjectures about the printing history of the *Transactions* and for providing me with additional background information; also Professor Dirk Siefkes for his help in acquiring Figures 4(c) and 11(c), and the Kunstbibliothek Berlin der Staatlichen Museen Preussischer Kulturbesitz for permission to publish them; and to André Jammes for permission to publish Figures 6 and 11(d).

**Bibliography.** The references below include several articles not referenced in the main text, namely a discussion of publishing at the American Institute of Physics [37]; some experiments in typesetting physics journals with the Bell Labs system [32], [7]; computer aids for technical magazine layout and editing, together with a brief proposal for a standard typesetting language [11]; reports about early computer programs for character generation and mathematics composition [23], [30], [34], [35], [40], [46]; a description of the mathematics a traditional printer needs to know [9]; three standard references on the typesetting of mathematics [14], [47], [50]; some fonts of type and special characters designed by the American Mathematical Society [39]; a recent and highly significant approach to mathematical definition of traditional type faces based on conic sections and on one-dimensional splines [15]; a proposal for a new way to control the spacing between letters based on somewhat mathematical principles [28]; and two purely mathematical papers inspired by typography [13], [21].

## References

1. A. V. Aho, S. C. Johnson, and J. D. Ullman, *Typesetting by* ACM *considered harmful*, Communications of the ACM **18** (1975), 740.

2. American Mathematical Society, *Development of the Photon for efficient mathematical composition*, Final report (May 10, 1965), National Science Foundation grant G-21913; NTIS number PB168627.

3. American Mathematical Society, *Development of computer aids for tape-control of photocomposing machines*, Report No. 2 (July 1967), *Extension of the system of preparing a computer-processed tape to include the setting of multiple line equations*, National Science Foundation grant GN-533; NTIS number PB175939.

4. American Mathematical Society, *Development of computer aids for tape-control of photocomposing machines*, Final report, Section B (August 1968), *A system for computer-processed tape composition to include the setting of multiple line equations*, National Science Foundation grant GN-533; NTIS number PB179418.

5. American Mathematical Society, *Development of computer aids for tape-control of photocomposing machines*, Final report, Section C (January 1969), *Implementation, hardware, and other systems*, National Science Foundation grant GN-533; NTIS number PB182088.

6. American Mathematical Society, *To complete the study of computer aids for tape-control of composing machines by developing an operating system*, Final report, no. AMATHS-CAIDS-71-0 (April 1971), National Science Foundation grant GN-690; NTIS number PB200892.

7. American Physical Society, *APS tests computer system for publising operations*, Physics Today **30**, 12 (December 1977), 75.

8. Donald M. Anderson, *Cresci and his capital alphabets*, Visible Language **4** (1971), 331–352.

9. J. Woodard Auble, *Arithmetic for printers*, second ed., Peoria, Ill., Bennett, 1954.

10. Michael P. Barnett, *Computer typesetting: Experiments and prospects*, Cambridge, Mass., M.I.T. Press, 1965.

11. Robert W. Bemer and A. Richard Shriver, *Integrating computer text processing with photocomposition*, IEEE Trans. on Prof. Commun. PC-16 (1973), 92–96. This article is reprinted with another typeface and page layout in Robert W. Bemer, *The role of a computer in the publication of a primary journal*, Proc. AFIPS Nat. Comput. Conf. 42, Part II (1973), M16-M20.

12. Peter J. Boehm, *Software and hardware considerations for a technical typesetting system*, IEEE Trans. on Prof. Commun. PC-19 (1976), 15–19.

13. J. A. Bondy, *The 'graph theory' of the Greek alphabet*, Graph Theory and Applications, Y. Alavi et al., eds., Berlin, Springer-Verlag, 1972, pp. 43–54.

14. Theodore William Chaundy, Percy Reginald Barrett, and Charles Batey, *The printing of mathematics*, Oxford, Oxford Univ. Press, 1954.

15. P. J. M. Coueignoux, *Generation of roman printed fonts*, Ph. D. thesis, Dept. of Electrical Engineering, M.I.T., June, 1975.

16. Giovanni Francesco Cresci Milanese, *Essemplare de piv sorti lettere*, Rome, 1560. Also edited and translated by Arthur Sidney Osley, London, 1968.

17. T. L. De Vinne, *The practice of typography: Modern Methods of book composition*, New York, Oswald, 1914.

18. Albrecht Dürer, *Underweysung der Messung mit dem Zirckel und Richtscheyt*, Nuremberg, 1525. An English translation of the section on alphabets has been published as Albrecht Dürer, *Of the just shaping of letters*, R. T. Nichol, trans., Dover, 1965.

19. Felice Feliciano Veronese, *Alphabetum romanum*, Giovanni Mardersteig, ed., Verona, Editiones Officinae Bodoni, 1960.

20. Frederic W. Goudy, *Typologia: Studies in type design and type making with comments on the invention of typography, the first types, legibility and fine printing*, Berkeley, Calif., Univ. of California Press, 1940.

21. F. Harary, *Typographs*, Visible Language **7** (1973), 199–208.

22. Roar Hauglid, Randi Asker, Helen Engelstad, and Gunvor Traetteberg, *Native art of Norway*, Oslo, Dreyer, 1965.

23. A. V. Hershey, *Calligraphy for computers*, NWL Report No. 2101, Dahlgren, Va., U. S. Naval Weapons Laboratory, August 1967; NTIS number AD662398.

24. André Jammes, *La réforme de la typographie royale sous Louis XIV*, Paris, Paul Jammes, 1961.

25. Paul E. Justus, *There is more to typesetting than setting type*, IEEE Trans. on Prof. Commun. PC-15 (1972), 13–16.

26. Alan C. Kay, *Microelectronics and the personal computer*, Scientific American **237**, 3, September 1977, 230–244.

27. Brian W. Kernighan and Lorinda L. Cherry, *A system for typesetting mathematics*, *Communications of the* ACM **18** (1975), 151–157.

28. David Kindersley, *Optical letter spacing for new printing systems*, London, Wynkyn de Worde Society, 1976.

29. Donald E. Knuth, *Tau Epsilon Chi, a system for technical text*, Stanford Computer Science report CS675, September, 1978. To be published by the American Mathematical Society.

30. Dorothy K. Korbuly, *A new approach to coding displayed mathematics for photocomposition*, IEEE Trans. on Prof. Commun. PC-18 (1975), 283–287.

31. E. H. Lee and G. E. Forsythe, *Variational study of nonlinear splines*, SIAM Rev. **15** (1973), 120–133.

32. M. E. Lesk and B. W. Kernighan, *Computer typesetting of technical journals on* UNIX, Computing Science Tech. Report 44, Murray Hill, N. J., Bell Laboratories, June, 1976.

33. Giovanni Mardersteig, *The alphabet of Francesco Torniello* (1517) *da Novara*, Verona, Officini Bodoni, 1971.

34. M. V. Mathews and Joan E. Miller, *Computer editing, typesetting, and image generation*, Proc. AFIPS Fall Joint Computer Conf. **27** (1965), 389–398.

35. M. V. Mathews, Carol Lochbaum and Judith A. Moss, *Three fonts of computer drawn letters*, Communications of the ACM **10** (1967), 627–630.

36. Even Mehlum, *Nonlinear splines*, Computer Aided Geometric Design, Robert E. Barnhill and Richard F. Riesenfeld, eds., New York, Academic Press, 1974, pp. 173–207.

37. A. W. Kenneth Metzner, *Multiple use and other benefits of computerized publishing*, IEEE Trans. on Prof. Commun. PC-18 (1975), 274–278.

38. Joseph Moxon, *Regulae trium ordinum literarum typographicarum, or the rules of the three orders of print letters: viz. the {roman, italick, english} capitals and small; Shewing how they are compounded of Geometrick Figures, and mostly made by Rule and Compass*, London, Joseph Moxon, 1676.

39. Phoebe J. Murdock, *New alphabets and symbols for typesetting mathematics*, Scholarly Publishing **8** (1976), 44–53. Reprinted in Notices Amer. Math. Soc. **24** (1977), 63–67.

40. Nicholas Negroponte, *Raster scan approaches to computer graphics*, Computers and Graphics **2** (1977), 179–193.

41. Wolfgang A. Ocker, *A program to hyphenate English words*, IEEE Trans. on Prof. Commun. PC-18 (1975), 78–84.

42. Luca Pacioli, *Divina Proportione, Opera a tutti glingegni perspicaci e curiosi necessaria Ove ciascun studioso di Philosophia, Propectiva, Pictura, Sculptura: Architecturo: Musice: altre Mathematice: suavissima: sottile: e admirable et doctrina consequira: e delectarassi: con varie questione de secretissima scientia* (Venice, 1509).

43. Giovanbattista Palatino Cittadino Romano, *Libro primo del le lettere maiuscole antiche romane* (unpublished), Berlin Kunstbibliothek, MS. OS5280. Some of the individual pages are dated 1543, 1546, 1549, 1574, or 1575. See James Wardrop, *Civis romanus sum: Giovanbattista Palatino and his circle*, Signature, n.s. 14 (1952), 3–39.

44. Paul A. Parisi, *Composition innovations of the American Society of Civil Engineers*, IEEE Trans. on Prof. Commun. PC-18 (1975), 244–273.

45. R. G. D. Richardson, *The twenty-ninth annual meeting of the Society*, Bull. Amer. Math. Soc. **29** (1923), 97–116. (See also vol. **28** (1922) pp. 234–235, 378 for comments on the special *Transactions* volume, and pp. 2–3 of vol. 28 for discussion of deficits due to increased cost of printing.)

46. Glenn E. Roudabush, Charles R. T. Bacon, R. Bruce Briggs, James A. Fierst, Dale W. Isner and Hiroshi A. Noguni, *The left hand of scholarship: Computer experiments with recorded text as a communication media*, Proc. AFIPS Fall Joint Computer Conf. **27** (1965), 399–411.

47. Ellen E. Swanson, *Mathematics into type*, Amer. Math. Soc., Providence, R. I., 1971.

48. Francesco Torniello, *Opera del modo de fare le littere mauiscuole antique*, Milan, Italy, 1517.

49. Geofroy Tory, *Champ fleury*, Paris, 1529. Also translated into English and annotated by George B. Ives, New York, Grolier Club, 1927.

50. Karel Wick, *Rules for typesetting mathematics*, translated by V. Boublik and M. Hejlová, The Hague, Mouton, 1965.

51. Hermann Zapf, *About alphabets: some marginal notes on type design*, Cambridge, Mass., M.I.T. Press, 1970.

DEPARTMENT OF COMPUTER SCIENCE, STANFORD UNIVERSITY. STANFORD, CALIFORNIA 94305