114

The Letter S

Donald E. Knuth

SEVERAL YEARS AGO when I began to look at the problem of designing suitable alphabets for use with modern printing equipment, I found that 25 of the letters were comparatively easy to deal with. The other letter was 'S'. For three days and nights I had a terrible time trying to understand how a proper 'S' could really be defined. The solution I finally came up with turned out to involve some interesting mathematics, and I believe that students of calculus and analytic geometry may enjoy looking into the question as I did. The purpose of this paper is to explain what I now consider to be the 'right' mathematics underlying printed S's, and also to give an example of the METAFONT language I have recently been developing. (A complete description of METAFONT, which is a computer system and language intended to aid in the design of letter shapes, appears in [3, part 3].

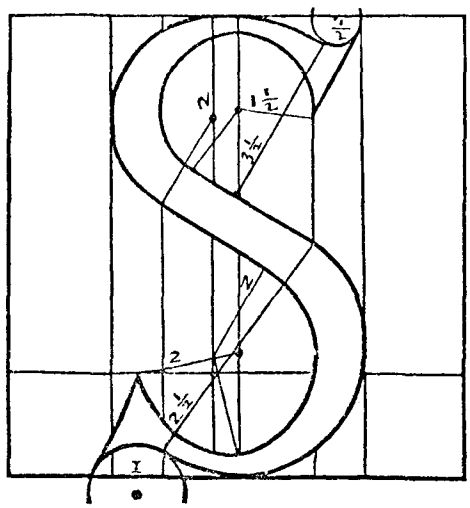
Before getting into a technical discussion, I should probably mention why I started worrying about such things in the first place. The central reason is that today's printing technology is essentially based on discrete mathematics and computer science, not on properties of metals or of movable type. The task of making a plate for a printed page is now essentially that of constructing a gigantic matrix of 0's and 1's, where the 0's specify white space and the 1's specify ink. I wanted the second edition of one of my books to look like the first edition, although the first edition had been typeset with the old hot-lead technology; and when I realized that this problem could be solved by using appropriate techniques of discrete mathematics and computer science, I couldn't resist trying to find my own solution.

Reference [2] explains more of the background of my work, and it also discusses the early history of mathematical approaches to type design. In particular, it illustrates how several people proposed to construct S's geometrically with ruler and compass during the sixteenth and seventeenth centuries.

Francesco Torniello published a geometric alphabet in 1517 that is typical of these early approaches. Let's look at his construction of an 'S' (cf. Fig. 1),

The preparation of this article was supported in part by National Science Foundation grants MCS-7723738 and IST-7921977, by Office of Naval Research grant N00014-76-C-0330, and by the IBM Corporation. The author gratefully acknowledges the help of Xerox Palo Alto Research Laboratory facilities for the preparation of several illustrations. All of the letters and symbols in this report were designed mathematically, using METAFONT.

.1., con quello tondo quale ha lo suo puncto de mezo fora del quadro, longe da la inferiore linea del quadro puncto mezo. Poi largo lo circino puncti .2., ponendo una puncta dove finisti la inferiore parte del .S. qual fu facta a drita linea, cioè longe da la linea del spacio da parte drita puncti .2., e altri



puncti .4. da la [linea] inferiore del quadro. L'altra puncta longe da quella del spacio da parte sinistra puncti .2. descenderai in tondo verso man drita tanto che giongi sopra la media linea. Poi con dicta largheza de circino ponendo l'una puncta dove al presente finisti, l'altra puncta longe da la linea del spacio da parte sinistra puncti .2., venendo dal dicto ultimo loco del .S. tanto che sia lontano da la inferiore linea del quadro puncti .2. Poi da questa ultima parte in tondo vengasi a drita linea a congiungere con lo inferiore tondo longe da la linea da parte sinistra del quadro puncti .1. e sette octavi; & sarà finita la littera .S., come apertamente si vede.

Fig. 1. Francesco Torniello's method of "squaring the S" in 1517. (This is page 45 of [4], reproduced by kind permission of Officina Bodoni in Verona, Italy.)

in order to get some feeling for the problems involved. Paraphrasing his words into modern mathematical terminology, we can state the method as follows:

> An 'S' is drawn in a 9×9 square that we can represent by Cartesian coordinates (x, y) for $0 \le x \le 9$ and $0 \le y \le 9$. We shall define fourteen points on the boundary of the letter, calling them (x_1, y_1), (x_2, y_2), ..., (x_{14}, y_{14}). Point 1 is $(4.5, 9)$, and a circular arc is drawn from this point with center at $(4.5, 5.5)$ and radius 3.5 ending at point 2 where $x_2 = 6$.

[Hence $y_2 = 5.5 + \sqrt{10} \approx 8.66$.] A small arc is drawn with center $(6.5, 9)$ and radius .5 from point $3 = (6.5, 8.5)$ to $(7, 9)$. A straight line is drawn from point $4 = (6, 7)$ to where it is tangent to this small arc; let us call this point 5. [We shall see below that point 5 has the coordinates $(6\frac{16}{17}, 8\frac{13}{17})$; it is interesting to speculate about whether Torniello would have been happy to know this.] Now an arc is drawn with center $(4, 7)$ and radius 2, from point $6 = (4, 9)$ down to point 7 where $x_7 = 3$ and $y_7 < 7$ [hence $y_7 = 7 - \sqrt{3} \approx 5.27$]. A straight line is drawn from point 7 to point $8 = (5, 4)$. An arc centered at $(4.5, 7\frac{1}{8})$ is now drawn from point 4 to point $9 = (3.5, 6)$, and a straight line continues from there to point $10 = (6, 4.5)$. A half-circle runs from this point to point $11 = (3, 0.5)$, with center $(4.5, 2.5)$ and radius 2.5. Another small circular arc is now drawn with center at $(2.5, y)$ and radius 1, from point 11 to point 12 where $x_{12} = 1\frac{7}{8}$ [hence $y = (1 - \sqrt{3})/2 \approx -.37$ and $y_{12} = (\sqrt{39} + 4 - 4\sqrt{3})/8 \approx 0.41$]. Circular arcs of radius 2 are drawn from point 8 to point 13 with the center x-coordinate equal to 4 and with $x_{13} = 4.5$ [hence the center is $(4, 4 - \sqrt{3} \approx 2.27)$ and $y_{13} = 4 - \sqrt{3} - \sqrt{3.75} \approx 0.33$], and from point 13 to point 14 with the center x-coordinate equal to 4.5 and with $y_{14} = 2$ [hence the center is $(4.5, 6 - \sqrt{3} - \sqrt{3.75} \approx 2.33)$ and $x_{14} = 4.5 - \sqrt{4 - (4 - \sqrt{3} - \sqrt{3.75})^2} \approx 2.53$.] Finally a straight line runs from point 14 to point 12.

The reader will find it interesting to take a piece of graph paper and carry out this vintage construction before proceeding further. Torniello's description was actually not so precise as this, and I have tried to make as much sense out of his words as possible; it seems that he had as much trouble with S's as I did, because his other letters are much more clearly defined. The main editorial revision I have made is to change the center of the arc between points 4 and 9 from Torniello's $(4.5, 7\frac{1}{6})$ to the nearby point $(4.5, 7\frac{1}{8})$, and to leave its radius unstated [he said the radius would be 1.5, but actually it is $\sqrt{145}/8$, a trifle higher], since $(4.5, 7\frac{1}{6})$ is not equidistant from points 4 and 9.

Note that the circular arc between points 10 and 11 is tangent to the baseline at $(4.5, 0)$ and it has a vertical tangent at point $(7, 2.5)$; this works out nicely because $3^2 + 4^2 = 5^2$, and I believe Torniello did know enough mathematics to make use of this pleasant coincidence in his design. He never stated exactly what curves should be used between points 1 and 6 or between 2 and 3; apparently a straight line segment should join 1 and 6, while the other curve is to be filled in with whatever looks right.

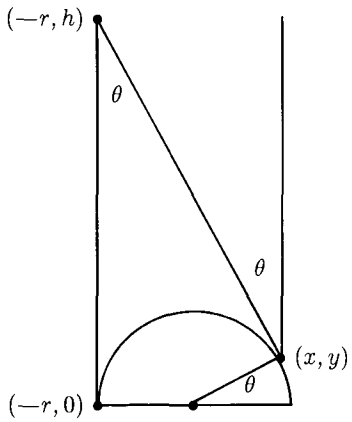


Fig. 2. A problem that arises in Torniello's construction: Find x and y, given r and h.

The calculation of point 5 suggests an elementary but instructive exercise in analytic geometry: *Given positive numbers h and r, find the point (x, y) in the upper right portion of a circle of radius r, centered at the origin, such that the straight line from $(-r, h)$ to (x, y) is tangent to the circle at (x, y).* (See Fig. 2.) We have $x^2 + y^2 = r^2$ and $y/x = \tan\theta = (x + r)/(h - y)$, hence $x^2 + rx + y^2 - yh = 0$ and $rx = hy - r^2$. This leads to the equation $0 = (hy - r^2)hy + r^2y(y - h) = rx(rx + r^2) + r^2y(y - h)$, hence $y(h^2y - hr^2 + r^2y - hr^2) = 0$ and we soon obtain the desired solution

$$x = \frac{h^2r - r^3}{h^2 + r^2}, \quad y = \frac{2hr^2}{h^2 + r^2}.$$

The solution is a rational function of h and r (i.e., no square roots are needed) because the other tangent point is $(-r, 0)$; this other point also satisfies the stated equations. René Descartes would surely have liked this demonstration of the power of his coordinate system.

Torniello's construction can be expressed without difficulty in the METAFONT language, a language that I have recently developed for stating definitions of character shapes in a form that is convenient for computer processing. Although ruler-and-compass methods do not really use very many of METAFONT's abilities, we can learn something about METAFONT by looking at this as a first example.

The key points of a particular design are specified in METAFONTese by writing equations for their x and y coordinates; then you can say "**draw** $i..j$" to draw a straight line from point i to point j. You can also say "**draw** $i\{\alpha,\beta\}..j\{\gamma,\delta\}$" to draw a curve from point i starting in the direction of the vector (α, β) and ending at point j in direction (γ, δ). This curve will be a circular arc if there is a circle passing through i and j in the stated directions, provided that the circular arc is at most a half-circle. Thus, Torniello's construction can be expressed with complete precision by the following METAFONT program:

$x_1 = 4.5u$; $y_1 = 9u$;
$x_2 = 6u$; $y_2 - 5.5u =$
sqrt$((3.5u)(3.5u) - (x_2 - 4.5u)(x_2 - 4.5u))$;
draw $1\{y_1 - 5.5u, 4.5u - x_1\}$..
$2\{y_2 - 5.5u, 4.5u - x_2\}$;
$x_3 = 6.5u$; $y_3 = 8.5u$;
$x_4 = 6u$; $y_4 = 7u$;
$x_5 = (6 + \frac{16}{17})u$; $y_5 = (8 + \frac{13}{17})u$;
draw $3\{9u - y_3, x_3 - 6.5u\}$..
$5\{9u - y_5, x_5 - 6.5u\}$;
draw 4..5;
$x_6 = 4u$; $y_6 = 9u$;
$x_7 = 3u$; $7u - y_7 =$
sqrt$((2u)(2u) - (x_7 - 4u)(x_7 - 4u))$;
draw $6\{7u - y_6, x_6 - 4u\}..7\{7u - y_7, x_7 - 4u\}$;
$x_8 = 5u$; $y_8 = 4u$; **draw** 7..8;
$x_9 = 3.5u$; $y_9 = 6u$;
$x_{15} = 4.5u$; $y_{15} - 7.125u =$
sqrt$((x_9 - 4.5u)(x_9 - 4.5u) +$
$(y_9 - 7.125u)(y_9 - 7.125u))$;
draw $4\{7.125u - y_4, x_4 - 4.5u\}..15..$
$9\{7.125u - y_9, x_9 - 4.5u\}$;
$x_{10} = 6u$; $y_{10} = 4.5u$; **draw** 9..10;
$x_{11} = 3u$; $y_{11} = .5u$;
draw $10\{y_{10} - 2.5u, 4.5u - x_{10}\}..$
$11\{y_{11} - 2.5u, 4.5u - x_{11}\}$;
$x_{16} = 2.5u$; $y_{11} - y_{16} =$
sqrt$(u{\cdot}u - (x_{11} - x_{16})(x_{11} - x_{16}))$;
$x_{12} = 1.875u$; $y_{12} - y_{16} =$
sqrt$(u{\cdot}u - (x_{12} - x_{16})(x_{12} - x_{16}))$;
draw $11\{y_{16} - y_{11}, x_{11} - x_{16}\}..$
$12\{y_{16} - y_{12}, x_{12} - x_{16}\}$;
$x_{13} = 4.5u$; $x_{17} = 4u$; $y_8 - y_{17} =$
sqrt$((2u)(2u) - (x_8 - x_{17})(x_8 - x_{17}))$;
$y_{17} - y_{13} =$
sqrt$((2u)(2u) - (x_{13} - x_{17})(x_{13} - x_{17}))$;
draw $8\{y_8 - y_{17}, x_{17} - x_8\}..$
$13\{y_{13} - y_{17}, x_{17} - x_{13}\}$;
$x_{18} = 4.5u$; $y_{18} - y_{13} =$
sqrt$((2u)(2u) - (x_{18} - x_{13})(x_{18} - x_{13}))$;
$y_{14} = 2u$; $x_{18} - x_{14} =$
sqrt$((2u)(2u) - (y_{18} - y_{14})(y_{18} - y_{14}))$;
draw $13\{y_{13} - y_{18}, x_{18} - x_{13}\}..$
$14\{y_{14} - y_{18}, x_{18} - x_{14}\}$;
draw 14..12.

Here "u" is an arbitrary unit of measure that can be used as a scale factor to control the overall size of the drawing. This program looks somewhat formidable at first glance, but it really is not hard to understand once you compare it to the informal English description given earlier. A few more points, labeled 15, 16, 17, and 18, have been introduced; point 15 coaxes METAFONT to draw a circular arc bigger than a semicircle, and the other three points are centers of arcs in the construction. The main fact used throughout is that a circular arc with center (x_k, y_k) that passes clockwise through point (x_i, y_i) is going in direction $\{y_i - y_k, x_k - x_i\}$, while if the arc is going counterclockwise its direction is $\{y_k - y_i, x_i - x_k\}$.

Fig. 3 shows what METAFONT draws from the above specifications. METAFONT will also complete

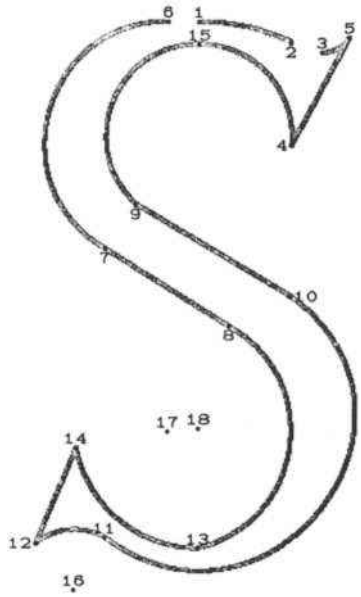


Fig. 3. The METAFONT program in the text will produce this rendition of Torniello's S.

the drawing with appropriate non-circular curves if we add the commands

draw 1..6;
draw $2\{y_2 - 5.5u, 4.5u - x_2\}..3\{9u - y_3, x_3 - 6.5u\}$.

These tangent directions match the tangents at which the new curves touch the old. If we ask METAFONT to fill in the space between these boundary curves, we obtain Fig. 4.

When the circular arc comes to point 7 from point 6 it is travelling in direction $\{7u - y_7, x_7 - 4u\} = \{\sqrt{3}u, -u\}$, but when it proceeds from point 7 in a straight line to point 8 it abruptly shifts to direction

Fig. 4. The curve of Fig. 3, completed and filled in.

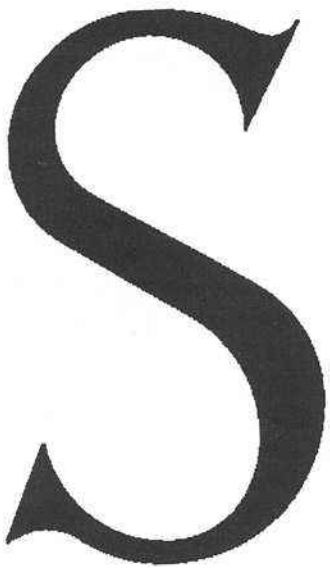


Fig. 5. A slight modification of Fig. 4 makes the curves smoother at the junction points.

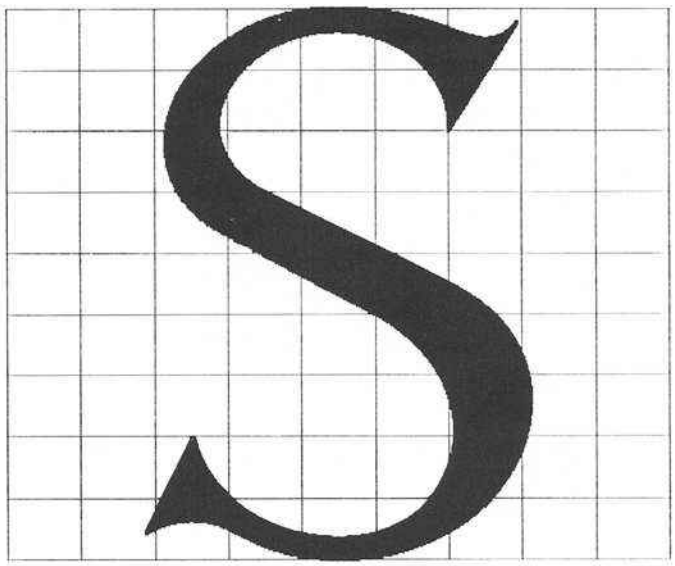


Fig. 6. When Fig. 5 is stretched 20% in the horizontal direction, we obtain this figure; the circles have become ellipses.

$\{x_8 - x_7, y_8 - y_7\} = \{2u, (\sqrt{3} - 3)u\}$. This discontinuity is only slightly noticeable in Fig. 4, but it is unsatisfactory from a mathematical standpoint. Similar discontinuities occur at points 8, 9, 10, and 13, the problems at points 9 and 13 being especially prominent; the illustration in Torniello's book had to be fudged slightly to hide these defects (which Torniello did not mention). Contemporary standards of accuracy were presumably not very stringent in the sixteenth century, but nowadays we do not want our computers to draw such crooked lines.

Since METAFONT has no special commitment to circular arcs, it will automatically make adjustments like Torniello's illustrator did if we just specify consistent directions at all of the key points. Fig. 5 shows the result if the tangents at points 7, 8, 9 and 10 are taken as the directions of the straight line segments and if the direction at point 13 is horizontal. Furthermore point 6 has been moved over to coincide with point 1, so that the unfortunate flat spot at the top is avoided. The curves touching these points are not circles any longer, but they are close enough to fool most people, and it seems unlikely that Torniello would have been offended by this approximation.

A Renaissance 'S' looks somewhat skinny to modern eyes. We can ask METAFONT to flesh it out by increasing all the x coordinates by 20% while leaving the y coordinates fixed; Fig. 6 shows the result. Note that this stretching turns circles into ellipses. Torniello would have had considerable difficulty trying to specify such a shape in terms of strictly circular arcs; we are reminded of the early astronomers who found it very cumbersome to use circles instead of ellipses as models of planetary orbits.

By studying this example we can get some idea of the problems involved in specifying a proper S shape. However, I was actually seeking the solution to a more general problem than the one Torniello faced: Instead of specifying only one particular 'S', I needed many different variations, including **bold face** versions that are much darker than the normal text. I discussed this recently with Alan Perlis, who pointed out that a central issue arising whenever we try to automate something properly is what he calls "the art of making constant things variable." In the case of letter design, we don't merely want to take a particular drawing and come up with some mathematics to describe it; we really want to find the principles *underlying* the drawing, so that we can generate infinitely many drawings (including the given one) as a function of appropriate parameters. My goal was to create entire alphabets that would depend on a dozen or two parameters in such a way that all the letters would vary in a compatible manner as the parameters would change.

After looking at these Renaissance constructions and a lot of modern S shapes, I came to the conclusion that the main stroke of the general S curve I sought would be analogous to the curve in Fig. 6: each boundary curve was to be an ellipse followed by a straight line followed by another ellipse. This led me to pose the following problem: *What ellipse has its topmost point at (x_t, y_t) and its leftmost point at (x_l, y_l) for some y_l, and is tangent to the straight line of slope σ that passes through (x_c, y_c), given the values of x_t, y_t, x_l, σ, x_c, and y_c?* (The ellipse in question is supposed to have the coordinate axes as its major and minor axes; in other words, it should have left-right symmetry.) The reason for my posing this problem should be fairly clear from

our previous discussion: We know a point that is supposed to be the top of the S curve, and we also know how far the curve should extend to the left; furthermore we have a straight line in mind that will form the middle link of the stroke.

The problem stated in the preceding paragraph is interesting to me for several reasons. In the first place, it has a nice answer (as we will see). In the second place, the answer does in fact lead to satisfactory S curves. In the third place, the answer isn't completely trivial; during a period of two years or so I came across this problem four different times and each time I was unable to find my notes about how to solve it, so I spent several hours deriving and rederiving the formulas whenever I needed them. Finally I decided to write this paper so that I wouldn't have to derive the answer again.

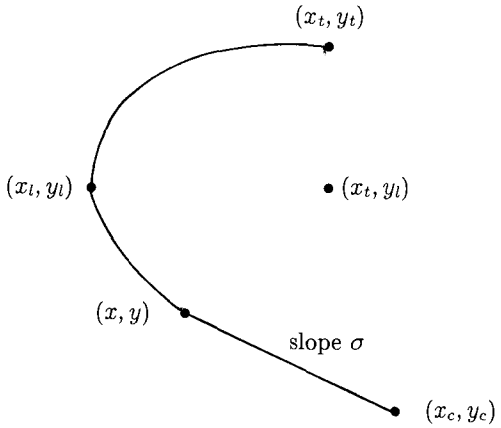


Fig. 7. Problem: Find x, y, and y_l when x_t, y_t, x_l, σ, x_c, and y_c are given.

The point (x_t, y_l) is the center of the ellipse we seek. Let (x, y) be the point where the desired ellipse is tangent to the line of slope σ through (x_c, y_c), as shown in Fig. 7. Our problem boils down to solving three equations in the three unknowns x, y, and y_l:

$$\begin{aligned} \left(\frac{x-x_t}{x_l-x_t}\right)^2+\left(\frac{y-y_l}{y_t-y_l}\right)^2 &= 1; \\ \frac{y_c-y}{x_c-x} &= \sigma; \\ -\left(\frac{y_t-y_l}{x_l-x_t}\right)^2\frac{x-x_t}{y-y_l} &= \sigma. \end{aligned} \qquad (*)$$

The first of these is the standard equation for an ellipse, and the second is the standard equation for a line; the third is obtained by differentiating the first,

$$2\,dx\,\frac{x-x_t}{(x_l-x_t)^2}+2\,dy\,\frac{y-y_l}{(y_t-y_l)^2}=0,$$

and setting dy/dx equal to σ.

Before attempting to solve equations $(*)$, I would like to introduce a notation that has turned out to be extremely useful in my work on mathematical font design: Let $\alpha[x, y]$ be an abbreviation for

$$x+\alpha(y-x),$$

which may be understood as "the fraction α of the way from x to y". Thus $0[x, y] = x$; $1[x, y] = y$; $\frac{1}{2}[x, y]$ is the midpoint between x and y; $\frac{3}{4}[x, y]$ is halfway between y and this midpoint; and $2[x, y]$ lies on the opposite side of y from x, at the same distance as y is from x. Identities like $\alpha[x, x] = x$ and $\alpha[x, y] = (1-\alpha)[y, x]$ are easily derived. When making some geometric construction it is common to refer to things like the point one third of the way from A to B; the notation $\frac{1}{3}[A, B]$ means just that.

One of the uses of this bracket notation is to find the intersection (x, y) of two given lines, where the lines go respectively from (x_1, y_1) to (x_2, y_2) and from (x_3, y_3) to (x_4, y_4). We can solve the intersection problem by noting that there is some number α such that

$$x=\alpha[x_1,x_2], \qquad y=\alpha[y_1,y_2]$$

and some number β such that

$$x=\beta[x_3,x_4], \qquad y=\beta[y_3,y_4].$$

These four simultaneous linear equations in x, y, α, β are easily solved; and in fact METAFONT will automatically solve simultaneous linear equations, so it is easy to compute the intersection of lines in METAFONT programs.

The bracket notation also applies to ellipses in an interesting way. We can write $x = \alpha[x_0, x_{\max}]$ and $y = \beta[y_0, y_{\max}]$ in the general equation

$$\left(\frac{x-x_0}{x_{\max}-x_0}\right)^2+\left(\frac{y-y_0}{y_{\max}-y_0}\right)^2=1,$$

reducing it to the much simpler equation

$$\alpha^2+\beta^2=1.$$

Returning to our problem of the ellipse, let us set

$$\begin{aligned} x &= \alpha[x_t, x_l], & y &= \beta[y_l, y_t], \\ X &= x-x_t, & Y &= y_l-y_t, \\ a &= x_l-x_t, & b &= (y_c-\sigma x_c)-(y_t-\sigma x_t). \end{aligned}$$

The three equations $(*)$ can now be rewritten as follows:

$$\begin{aligned} \alpha^2+\beta^2 &= 1; \\ b+\sigma X &= (1-\beta)Y; \\ aY &= a\sigma\beta; \\ X &= a\alpha. \end{aligned} \qquad (**)$$

This gives us four equations in the four unknowns (α, β, X, Y), so it may seem that we have taken a step

backwards; but the equations are much simpler in form. We can eliminate a to reduce back to three unknowns:

$$X^2 + a^2\beta^2 = a^2; \tag{1}$$
$$b + \sigma X = (1-\beta)Y; \tag{2}$$
$$XY = a^2\sigma\beta. \tag{3}$$

Multiplying (3) by $(1-\beta)$ and applying (2) now leads to

$$X(b+\sigma X) = a^2\sigma\beta(1-\beta),$$

and this miraculously combines with (1) to yield

$$bX = a^2\sigma(\beta - 1). \tag{4}$$

It follows that $\big(a^2\sigma(\beta-1)\big)^2 + a^2b^2\beta^2 = a^2b^2$, i.e.,

$$a^2(\beta-1)\big(a^2\sigma^2(\beta-1) + b^2(\beta+1)\big) = 0. \tag{5}$$

If $a = 0$, our equations become degenerate, with infinitely many solutions $(X,Y) = \big(0, b/(1-\beta)\big)$ for $-1 \le \beta < 1$. If $b = 0$, another degenerate situation occurs, with no solution possible unless $a\sigma = 0$, in which case there are infinitely many solutions with Y arbitrary and $(X,\alpha,\beta) = (0,0,1)$. Otherwise it is not difficult to see that $\beta \neq 1$, so (5) determines the value of β uniquely, and we can use this with (4) to determine the full solution:

$$\begin{aligned} \alpha &= -2ab\sigma/(a^2\sigma^2 + b^2); \\ \beta &= (a^2\sigma^2 - b^2)/(a^2\sigma^2 + b^2); \\ X &= -2a^2b\sigma/(a^2\sigma^2 + b^2); \\ Y &= (b^2 - a^2\sigma^2)/2b. \end{aligned} \tag{6}$$

I was surprised to find that the simultaneous quadratic equations (**) have purely rational expressions as their roots. There is a curious similarity between this solution and the answer to the problem in Fig. 2.

Translating (6) back into the notation of the original problem statement (Fig. 7), let (x_t, y_m) be on the line of slope σ through (x_c, y_c), so that $y_m = y_c + \sigma(x_t - x_c)$. Then the unique solution is

$$\begin{aligned} x &= x_t + \frac{2\sigma(x_l - x_t)^2(y_t - y_m)}{\sigma^2(x_l - x_t)^2 + (y_t - y_m)^2}, \\ y &= y_m + \frac{2\sigma^2(x_l - x_t)^2(y_t - y_m)}{\sigma^2(x_l - x_t)^2 + (y_t - y_m)^2}, \\ y_l &= y_t - \frac{(y_t - y_m)^2 - \sigma^2(x_l - x_t)^2}{2(y_t - y_m)}, \end{aligned} \tag{7}$$

except in the degenerate cases $x_l = x_t$ or $y_m = y_t$.

Incidentally, I tried the automatic equation-solving feature of the MACSYMA computer algebra system [5,7] on this problem, in order to get some idea of how long it will be before mathematicians will be replaced by computers when such calculations are required. MACSYMA correctly found the solution (X, Y, β) for equations (1), (2), (3) in about 17 seconds, except that it said nothing about the degenerate solutions that occur when $ab = 0$. The time required for MACSYMA to solve the system of four equations (**) was essentially the same as to deal with (1), (2), (3). But when I asked MACSYMA to solve the three original equations (*) for x, y, and y_l, the computer's memory capacity was exceeded after about a minute and twenty seconds, even when I simplified (*) by replacing (x_c, y_c) by (x_t, y_m). Thus, I was reassured to find that the equations (*) aren't completely trivial and that the conversion to (**) was an important step.

The above solution to the ellipse problem leads immediately to the desired S curves, since we can fill in the space between an ellipse-and-straight-line arc that runs from (x_t, y_t) to $(x_l^{(1)}, y_l^{(1)})$ to $(x^{(1)}, y^{(1)})$ to $(x_c, y_c^{(1)})$ and another that runs from (x_t, y_t) to $(x_l^{(2)}, y_l^{(2)})$ to $(x^{(2)}, y^{(2)})$ to $(x_c, y_c^{(2)})$, where the distance between $x_l^{(1)}$ and $x_l^{(2)}$ is governed by the desired thickness of the stroke at the left and the distance between $y_c^{(1)}$ and $y_c^{(2)}$ is governed by the desired thickness of the stroke at the center. (See Fig. 8. The actual S curve is drawn with a circular pen of small but positive radius whose *center* traces the curves shown, so the actual boundary is not a perfect ellipse.) The bottom right part of the S is, of course, handled in the same way as the upper left part.

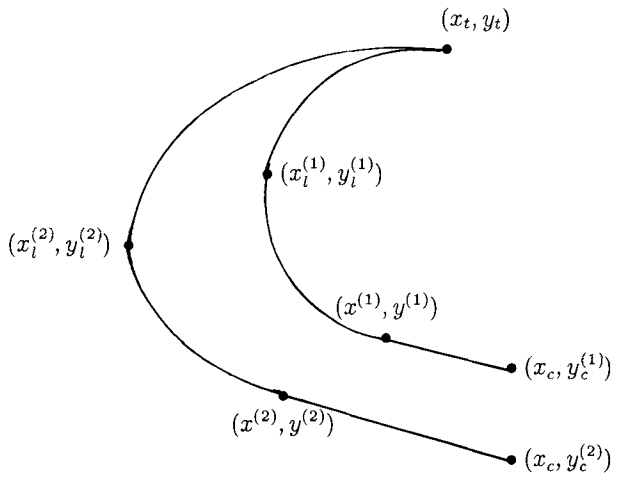


Fig. 8. A good S is obtained by drawing two partial ellipses according to the method of Fig. 7, then filling in the space between them, using a pen whose diameter is the width of the "hairlines" of the desired letters.

SSSSSS

Fig. 9. Different possibilities can be explored by varying the parameters. Here the slope is changing, but other characteristics are held fixed; the respective slopes are $\frac{2}{5}$, $\frac{1}{2}$, $\frac{2}{3}$, 1, $\frac{3}{2}$, 2, and $\frac{5}{2}$ times the "correct" slope in the middle.

Fig. 9 shows various S curves drawn by this method when the slope σ varies but the other specifications stay the same. Fig. 10 shows an S that has the same slope as the middle one of Fig. 9, but the curve is wider when it is travelling vertically at the upper left and the lower right. One of the chief advantages of a mathematical, parameterized approach is that it is easy to make lots of experiments until you find the setting of parameters that you like best. A METAFONT program that would draw the S's in Figs. 9 and 10, depending on appropriate parameters, appears in the appendix below.

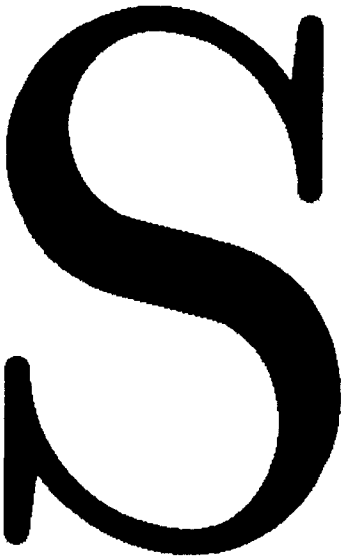


Fig. 10. The main stroke of this S is wider at the upper left and lower right, but otherwise it was drawn to the specifications of the middle S in Fig. 9.

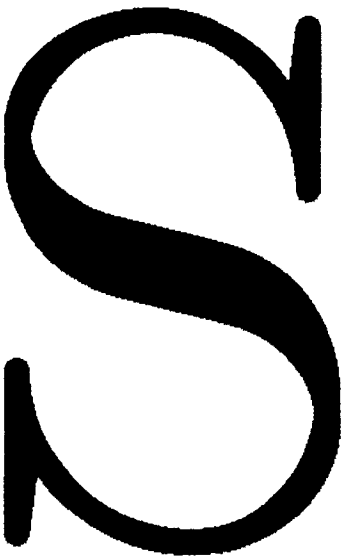


Fig. 11. Disastrous effects can occur if there isn't enough width at the upper left and lower right.

I happily made S's with this method for more than two years, but one day I decided to ask METAFONT to draw a great big letter S and the resulting shape was unexpectedly ugly. Looking back at some of the other supposedly nice S's drawn previously, I started to notice an occasional defect that was comparatively innocuous at the small scales I had been working with. This defect became painfully apparent when everything was enlarged, so I realized that I still hadn't gotten to the end of the story.

Fig. 11 illustrates this new difficulty in a somewhat extreme form. In terms of the notation of Fig. 8, I had not placed x_l sufficiently far to the right of x_l, so the two ellipses through $(x_l^{(1)}, y_l^{(1)})$ and $(x_l^{(2)}, y_l^{(2)})$ actually *crossed* each other. This made the supposed inner boundary switch over and become the outer boundary and vice versa, a distinctly unpleasant result since I was not intending to have such a calligraphic effect in this case.

The problem of Fig. 11 goes away if $x_l^{(1)}$ is sufficiently large, but of course it is desirable to know what the permissible values are. We are led to a third (and final) problem concerning ellipses: *What is a necessary and sufficient condition that the elliptical arc from $(x_l^{(2)}, y_l^{(2)})$ to (x_t, y_t) stays above the elliptical arc from $(x_l^{(1)}, y_l^{(1)})$ to (x_t, y_t)?* (We are assuming that $x_l^{(2)} < x_l^{(1)} < x_t$ and $y_l^{(2)} < y_l^{(1)} < y_t$, and that both ellipses have left/right symmetry as before.) It turns out that the answer to this problem can be expressed quite simply: the curves fail to cross if and only if

$$\frac{y_t - y_l^{(1)}}{(x_t - x_l^{(1)})^2} \geq \frac{y_t - y_l^{(2)}}{(x_t - x_l^{(2)})^2}. \tag{8}$$

My first attempt to find the right condition got bogged down in a notational mess, but finally I hit on the following fairly simple solution to this problem: Let $a = x_t - x_l^{(1)}$, $b = y_t - y_l^{(1)}$, $A = x_t - x_l^{(2)}$, and $B = y_t - y_l^{(2)}$. By turning the curves upside down, we want the function $b - b\sqrt{1-(x-a)^2}$ (which describes the bottom right quarter of an elliptical arc from $(0,0)$ to (a,b)) to be less than or equal to the analogous function $B - B\sqrt{1-(x-A)^2}$, whenever $|x| < a$. Expanding in power series we have

$$b - b\sqrt{1-(x/a)^2} = b\left(\frac{x^2}{2a^2} + \frac{x^4}{8a^4} + \cdots + \binom{1/2}{k}(-1)^{k+1}\frac{x^{2k}}{a^{2k}} + \cdots\right),$$

where

$$\binom{1/2}{k}(-1)^{k+1} = \frac{(2k-2)!}{2^{2k-1}\,k!\,(k-1)!}$$

is positive for all $k > 0$, and the power series converges for $|x| < a$. If $b/a^2 < B/A^2$, the analogous power

Fig. 12. Varying thicknesses of the middle stroke lead to these S's, where the width at upper left and lower right has been chosen to be as small as possible without the "crossover" problem of Fig. 11.

series

$$B - B\sqrt{1-(x/A)^2} = B\left(\frac{x^2}{2A^2}\frac{x^4}{8a^4} + \cdots + \binom{1/2}{k}(-1)^{k+1}\frac{x^{2k}}{A^{2k}} + \cdots\right),$$

will grow faster for small x and the two curves will cross. But if $b/a^2 \geq B/A^2$, we will have $b/a^{2k} \geq B/a^{2k}$ for all $k > 0$, so every term of the first power series dominates every term of the second. Q.E.D.

According to the theory worked out earlier, we have

$$\frac{y_t - y_l}{(x_t - x_l)^2} = \frac{y_t - y_m}{2(x_t - x_l)^2} - \frac{\sigma^2}{2(y_t - y_m)}. \qquad (9)$$

Thus we can ensure that $(y_t - y_l^{(1)})/(x_t - x_l^{(1)})^2$ is actually equal to $(y_t - y_l^{(2)})/(x_t - x_l^{(2)})^2$ by starting with desired values of x_t, y_t, $x_l^{(2)}$, $y_m^{(1)}$, and $y_m^{(2)}$: first $y_l^{(2)}$ is determined, then $x_l^{(1)}$, and finally $y_l^{(1)}$.

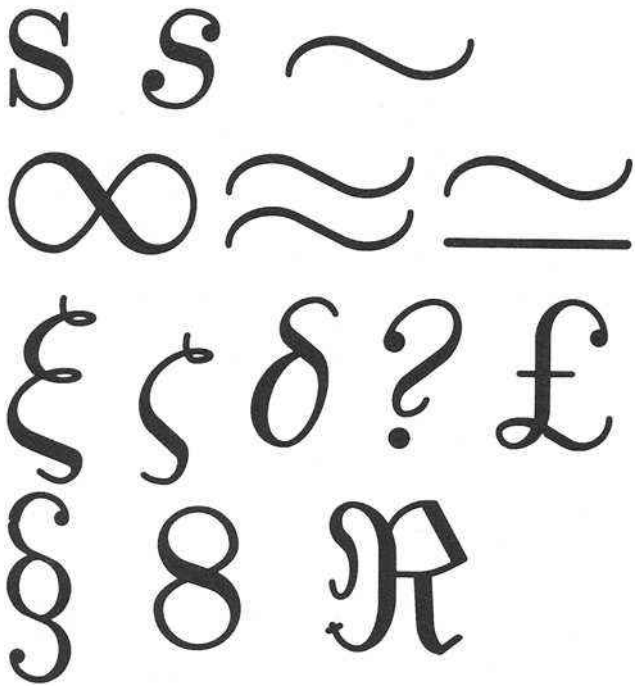


Fig. 13. The method used to draw an S stroke also is used as a subroutine that draws parts of many other characters, including those shown here.

After learning how to draw an S with mathematical precision, I found that the same ideas apply to many other symbols needed in a complete system of fonts for mathematics. In fact, all of the characters in Fig. 13 use the same METAFONT subroutine that I first developed for the letter S (or the dual subroutine obtained by interchanging x and y coordinates). Without the theory developed in this paper, I would either have had to abandon my goal of defining books in a mathematical way or I would have had to stop using all of these characters.

Of course, this is only a first step; the letters I have designed are far from optimal, and dozens of future experiments suggest themselves. My current dream is that the next several years will see mathematicians teaming up with experienced type designers to create truly beautiful new fonts. This will surely be one of the most visible applications of mathematics!

Let me close by asking a question of the reader. Ellipses have been studied for thousands of years, so it is reasonable to assume that all of their interesting properties were discovered long ago. Yet my experience is that when mathematics is applied to a new field, new 'purely mathematical' questions are often raised that enrich mathematics itself. So I am most curious to know: Have the questions that I encountered while trying to draw S-like ellipses been studied before, perhaps in some other disguise? Or did the new application of mathematics to typography lead to fresh insights about even such a well-studied object as a rectilinear ellipse?

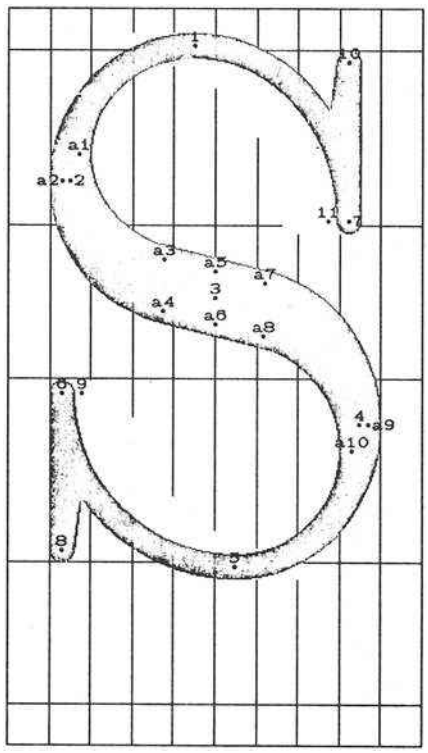


Fig. 14. The labeled points in this S correspond to the numbers specified by the METAFONT routine in the appendix.

Appendix

The METAFONT code below will draw the S shown in Fig. 14 (and infinitely many others) when the following parameters have been specified:

h, height of the character;
o, "overshoot" of curved lines at top and bottom;
u, one tenth of the character width;
w_0, size of circular pen used in drawing lines;
w_4, width of triangular serifs before erasing;
w_8, thickness of S stroke in the middle;
w_9, thickness at the upper left and lower right.

The vertical lines in Fig. 14 are u steps apart. The program uses "**lpen#**" and "**rpen#**" to erase unwanted ink that lies to the left and right of a specified path; the effect of such erasure is visible in the illustration, since portions of the guidelines have been erased.

```
subroutine scomp(index i)                    % starting point
        (index p)      % turning point (y_p to be defined)
        (index j)      % transition point (to be defined)
        (index k)                            % ending point
        (var s):                             % ending slope
% This subroutine computes y_p, x_j, and y_j so that
% y_k - y_j = s.(x_k - x_j) and so that the following curve
% is consistent with an ellipse:
% i{x_p - x_i,0}..p{0,y_p - y_i}..j{x_k - x_p,s.(x_k - x_p)}.
y_k - y_j = s(x_k - x_j);
new a,b;  a = s(x_p - x_i);  b = y_k - y_i - s(x_k - x_i);
x_j - x_i = -2a.b(x_p - x_i)/(a.a + b.b);
y_p - y_i = .5(b.b - a.a)/b.
```

```
subroutine sdraw(index i)                    % starting point
        (index p) % upper turning point (y_p to be defined)
        (index k)                            % middle point
        (index q) % lower turning point (y_q to be defined)
        (index j)                            % ending point
        (index a)  % effective pen width at turning points
        (index b)  % effective pen height at middle point
        (var s):                             % slope at middle point
cpen;  top_6 y_5 = top_b y_k;  bot_6 y_6 = bot_b y_k;
x_5 = x_6 = x_k;
rt_a x_p = rt_0 x_1;  lft_a x_p = lft_0 x_2;
rt_a x_q = rt_0 x_9;  lft_a x_q = lft_0 x_10;
y_2 = y_p;  y_9 = y_q;
call scomp(i,1,3,5,s);              % compute y_1 and point 3
call scomp(i,2,4,6,s);              % compute y_2 and point 4
call scomp(j,9,7,5,s);              % compute y_9 and point 7
call scomp(j,10,8,6,s);             % compute y_10 and point 8
hpen;  w_0 ddraw i{x_1 - x_i,0}..1{0,y_1 - y_i}..
        3{x_q - x_p,s(x_q - x_p)}..7{x_q - x_p,s(x_q - x_p)}..
        9{0,y_j - y_9}..j{x_j - x_9,0},
    i{x_2 - x_i,0}..2{0,y_2 - y_i}..
        4{x_q - x_p,s(x_q - x_p)}..8{x_q - x_p,s(x_q - x_p)}..
        10{0,y_j - y_10}..j{x_j - x_10,0}.   % the s-curve
"The letter S";
hpen;  top_0 y_1 = round(h + o);  bot_0 y_5 = -o;
x_3 = 5u;  y_3 = .52h;
lft_8 x_2 = round u;  rt_8 x_4 = round 9u;
x_1 = 4.5u;  x_5 = 5.5u;
lft_0 x_6 = round u;  rt_0 x_7 = round 8.5u;
y_6 = good_0 (1/3)h - 1;  y_7 = good_0 (2/3)h + 1;
bot_0 y_8 = 0;  y_9 = y_6;  x_8 = x_6;  rt_4 x_6 = rt_0 x_9;
top_0 y_10 = h;  y_11 = y_7;  x_10 = x_7;  lft_4 x_7 = lft_0 x_11;
w_0 ddraw 6..8,9..8;                          % lower serif
ddraw 7..10,11..10;                           % upper serif
rpen#;  w_4 draw 6{0,-1}..5{1,0};             % erase excess
lpen#;  w_4 draw 7{0,1}..1{-1,0};                    % ditto
hpen;  w_0 draw 6{0,-1}..5{1,0};              % lower left stroke
draw 7{0,1}..1{-1,0};                         % upper right stroke
call `a sdraw(1,2,3,4,5,8,9,-h/(50u)).        % middle stroke
```

References

1. Richard J. Fateman, *Essays in Algebraic Simplification*, Ph.D. thesis, Harvard University, April 1971; also MAC TR–95, April 1972. Available from MIT Laboratory for Computer Science, Publications, Room 112, 545 Technology Square, Cambridge MA 02139.
2. Donald E. Knuth, Mathematical Typography, *Bull. Amer. Math. Soc.* (new series) **1** (1979), 337–372. Reprinted with corrections as part 1 of [3].
3. Donald E. Knuth, *TEX and METAFONT: New Directions in Typesetting* (Providence, R.I.: American Mathematical Society, and Bedford, Mass.: Digital Press, 1979).
4. Giovanni Mardersteig, *The alphabet of Francesco Torniello (1517) da Novara* (Verona: Officina Bodoni, 1971).
5. The Mathlab Group, *MACSYMA Reference Manual*, version nine, 1977. Available from MIT Laboratory for Computer Science, Mathlab Group, Room 828, 545 Technology Square, Cambridge MA 02139. The original design and implementation of MACSYMA's "SOLVE" operator was due to R. J. Fateman, and it is briefly described in §3.6 of [1].
6. H. W. Mergler and P. M. Vargo, "One approach to computer-assisted letter design," *J. Typographic Research* **2** (1968), 299–322. [This paper describes the first computer system for drawing parameterized letters; for reasons that are now clear, the authors were unable to obtain a satisfactory 'S'!]
7. Joel Moses, MACSYMA—The Fifth Year, *Proc. EUROSAM '74*, Royal Inst. of Tech., Stockholm; *SIGSAM Bulletin* **8**,3 (Association for Computing Machinery, 1974), 105–110.

Department of Computer Science
Stanford University
Stanford, California, USA