The Grammar of Mathematical Symbolism

Fritz SCHWEIGER

IFFB Didaktik und LehrerInnenbildung, Universität Salzburg, Hellbrunnerstr. 34, A-5020 Salzburg

fritz.schweiger@sbg.ac.at

Abstract

The appearance of symbols is quite typical for mathematical texts. The use of symbols follows several rules which in most cases are not taught in an explicit manner but which are important to improve aspects of communication and cognition. The use of calculators, computer algebra, and word processors can the awareness of their functionality increase. Many of these rules are rooted in history and follow general semiotic principles.

1 Introduction

"What ideas do you connect with mathematics?" This question can provoke different answers but it will not come as a great surprise if you hear "$a^2 + b^2 = c^2$" or "Yes, I remember x and y". For many people mathematics has something to do with symbols and characters of a more or less dark meaning. Clearly, in other sciences you will also find symbols and formulas. Think of physics with the famous relation $E = mc^2$ or of chemistry with the well known H_2O. The development of symbolic systems is part of the history of mathematics and it can be shown that the development of apt notations was influential for the progress of mathematical thinking. We refer to Tropfke 1980 and Gericke 1984.

Mathematics uses language which is a subsystem of natural language enriched with peculiar signs and concepts (the so-called mathematical register, Halliday 1974; we refer also to Davis & Hunting 1990 and Maier & Schweiger 1999). A textbook can be written in English, German or Turkish but the employed symbols are similar around the world. The mathematical language is a *tool for doing mathematics* and a *medium of communication*. Mathematical contents are communicated with the use of the mathematical register but this language (think of written symbols and diagrams!) is a working medium as well. Mathematical symbols refer to notions but working with these symbols is part of mathematical activity. This is evident when looking at various calculations in written form or the solution of equations. In a manner similar to the study of natural languages one can distinguish between syntax and semantics of the mathematical language although the division line cannot be drawn sharply. In contrast to natural languages clearly phonology is not an important part because the mathematical register is (almost) a subsystem of a given natural language.

An important part of mathematics education is to teach a suitable knowledge of mathematical symbolisation. It is important to persuade students that a symbolic language is an indispensable help. Signs and symbols should be seen as an important help to understand mathematics and not as a barrier.

2 The system of mathematical symbols

As already mentioned mathematical symbols have been developed during a long historical process (see Tropfke 1980, Gericke 1984, Menninger 1979). The aim of these considerations is not to sketch the historical development but to analyse the implicit rules which govern the process leading to the 'mathematical pidgin' as we could call this system. Basically the relation between a symbol or sign and its meaning is arbitrary. A dog does not bear any sign that he is called *dog* in English or *köpek* in Turkish. But the need to communicate (and to work with the symbols) is a certain constraint.

The choice of symbols is regulated by at least three parameters: Tradition, communicability, and aspects of learnability. It is clearly tradition if an unknown number or a variable is denoted by the letter x. The ease of communication was a driving force in accepting the standard notation Θ for the set of rational numbers. The use of the arrow $\rightarrow$ for a map also bears the aspect of iconicity. From the viewpoint of learning the use of the first letter as r for radius or as A for area (clearly this aspect is dependent on the language of communication) can be recommended. On the other hand it could be important to avoid polysemy. Therefore in geometry one can use π for the circular number but then one must not use π to denote a projection. Some restrictions can be seen by international regulations as formulated by the International Organization for Standardization (ISO) and their national partners (http://www.iso.org/). These recommendations are not free from strange ideas such as the use of N for the set of natural numbers including 0. Clearly, the number 0 cannot be seen as a natural number because everyone counts 1, 2, 3, ... This looks a fossil from the exuberant use of set theory in mathematics education since 0 is the cardinal number of the empty set. A further restriction is the availability of characters and symbols on the computer. Some differentiating features like bold face cannot be used for handwriting.

Various classifications for mathematical symbols have been proposed. One may distinguish *visual* (or *iconic*) *symbols* and *algebraic* (or *verbal*) *symbols*, e.g. the sign Δ for a triangle in contrast to the letter x for a variable. But the use of Δ for the Laplace operator is just algebraic! This is again connected with the development of the mathematical notation. In early mathematical texts almost everything was expressed by whole words. Then a kind of syncopation (very often the use of the first letter of the word which denoted the concept) took place. One can show some nice cases in the development of this mathematical pidgin. The use of F for a closed set goes back to the French word *fermé* (= closed) and the use of G for an open set is related to the German word *Gebiet* (= domain; within topology the word is now reserved for a connected open set). Sometimes the meaning as well as the shape was changed. The standard symbol ∞ for infinity is a modified version of the Roman symbol M for 1 000 (in fact the use of M which is the first letter of Latin *mille* = 1 000 seems to be a later invention). The last stage is the more or less free use of symbols. In mathematical texts this assignment is signalled by phrases like 'We denote ...' or 'Let g be a straight line ...' In the German language this would be very appropriate since a straight line is *Gerade* (in Bahasa Indonesia it is *garislurus*).

It is also possible to differentiate between symbols, which denote the given data, and symbols, which refer to *activities*. In the phrase $25 \div 5$ the numbers refer to given data but the sign $\div$ signifies the activity (in this case the division) to be executed.

Another distinction can be made between symbols, which denote constants, and symbols, which denote *variables*. In a given text constants refer to the same concept and may be seen as the nouns of mathematical language. Variables are similar to pronouns. In a given text they can refer to different concepts. In the equation $x^2 + x - 1 = 0$ the letter x denotes a number which has to be found. In the formula $\int_0^1 2x \, \mathrm{d}x = 1$ the letter x means a so-called bound variable.

The system of mathematical symbols can be seen as an extension of a writing system. Alphabetic writing systems usually follow spoken language in their linear sequence of signs. The ideal is a writing system with a one-to-one correspondence of phonemes and graphemes. But most writing systems deviate in some way from this ideal. The writing of the English language is very deviant, e.g. the digraph *gh* can be spoken as an *f* in the word *laugh* but its appearance in the word *night* is due to an older pronunciation. In mathematics linearly ordered sequences and planar complex diagrams are used. The Chinese writing uses planar symbols as the carrier of meaning but their order follows spoken language. In some way between one should mention syllabic alphabets. The development of the world's writing systems is a very interesting part of cultural history (Haarmann 1991, Daniels & Bright 1996) and some of the strategies used in these systems are also used in mathematical symbolisation.

One should keep in mind that the correct reading of mathematical symbols is an achievement of its own. The context can be important. The symbol a_{11} seen as an element of a matrix is spoken as "a-one-one" but as a member of a sequence it could be "a-eleven". The correct reading of $\dfrac{\partial^2 f}{\partial x^2}$ also has to be learned. The sequence of symbols can follow the wording (in a given language!): $\sqrt{5}$ "square root of 5", a^2 as "a-square", $3 + 4 = 7$, "three plus four is seven", $\dfrac{4}{3}$ "four thirds" (to be read from above), 3^4 "three to the power of four" (to be read from left to right), and $\left(\begin{array}{c} n \\ 2 \end{array} \right)$ "n over two" (the brackets are read as "over"). The expression $\displaystyle\int_0^1 x^2 \, \mathrm{d}x$ is even more difficult to word correctly. The sequence of symbols can be different if one uses a hand calculator or a CAS.

Some symbols are pronounced according their semantic meaning: $a = b$ is spoken as "a is equal to b" but $a * b$ very often can be worded as "a star b" with the meaning of an algebraic operation. Letters normally are worded with their names: x is spoken "iks" but the letter has the meaning of a variable or unknown quantity. The correct wording of symbols can cause additional difficulties if one teaches or learns mathematics in a foreign language.

2.1 The origin of symbols

Mathematical symbols originate from various sources. There are the signs for numbers including several auxiliary symbols (decimal points, fraction bars and so on). The various alphabets build a great resource. This is the Latin alphabet, but also the Greek alphabet. The Hebrew letter ℵ (aleph) is used in set theory; the Cyrillic alphabet contributed the letter Ш (sha) for the Shafarevich group in algebraic geometry. Some symbols go back to letters but have been modified: the root symbol $\surd$ from Latin *radix* 'root', the symbol ∂ (mostly used for partial derivatives and the boundary operator) from *derivatio* 'derivation' or the integral sign $\displaystyle\int$ from Latin *summa* 'sum'.

There are a great number of special symbols which can be grouped together by similarity of form and meaning, for example the symbols for algebraic operations $+, *, \times, \circ$ or the symbols for symmetric relations (i.e. symbols denoting a kind of equality) $=, \sim, \equiv, \approx$.

Auxiliary symbols which are used as diacritic signs are a special class. Examples are strokes, stars, macrons $a', a*, \hat{a}$.

2.2 The formation of symbols

As just mentioned, the addition of other signs forms new symbols. In a more systematic way one can think of the following devices.

- *Numbers or letters in a lower position* to distinguish different objects: a_1, x_{23}, y_n.

- *Use of diacritic signs*: $x', \tilde{x}, \vec{x}$. This strategy is very old. In one ancient Greek system the letter which used ε for 5 but $_/\varepsilon$ for 5 000. This strategy is widespread in writing. In Turkish ş stands for a fricative sound like *sh* in shoe and contrasts with plain *s*. The similar distinction can be found in Arabic shin ش as contrasted with the letter sin س.
- *Letters or symbols in a higher position*: a^5, x^n, r^{-2}.
- *Juxtaposition*: $28, 2x^2, 3\frac{1}{2}$. These examples show that juxtaposition is open to different interpretations: $28 = 20 + 8$, $2x^2 = 2 \times x^2$, but $3\frac{1}{2} = 3 + \frac{1}{2}$ (and not $3 \times \frac{1}{2}$).
- *Planar symbols*: $\frac{3}{4}$, $\sqrt{c}$, $\sqrt[5]{x}$, $\sum_{i=1}^{\infty} \frac{1}{2^i}$, $|d|$, $\|y\|$, $\begin{vmatrix} -2 & 1{,}5 \\ 6{,}2 & -4 \end{vmatrix}$.

Symbols and chains of symbols have different meanings according to:

- *Order*: 17 is different from 71.
- *Position*: 23 is different from 2^3.
- *Size*: Indices and exponents are normally smaller in size. The symbol $\cap$ denotes the binary operation 'intersection' but the bigger symbol $\bigcap$ is used for the intersection of an arbitrary number of sets.
- *Shape*: The difference in *shape* distinguishes the types of brackets ( ), [ ], and { }. Here again this difference can be important as in the following example: In number theory $[x]$ denotes the integral part of x but $\{x\}$ means the fractional part of x. In the theory of Lie algebras $[x, y]$ is used for the binary operation. The use of { } in set theory is conventional. The equation $(3x + 5) - 2(x - 1) = 12$ is just more usual than $[3x + 5] - 2[x - 1] = 12$. There is a great difference in meaning between | | and ( ) as can be seen from examples like $|a + b| \leq |a| + |b|$ and $(a + b)c = ab + ac$ or $\begin{vmatrix} a & \\ c & d \end{vmatrix}$ (determinant) and $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ (matrix).
- *Orientation*: $\cap$ has different meaning from $\cup$, $\supseteq$ is different from $\subseteq$. To my knowledge only some syllabic alphabets for native languages of Canada use a similar device systematically. We give two examples from Inuktitut:

$$\begin{array}{ccc} \triangleleft a & \Delta i & \triangleright u \\ < pa \wedge pi >, & & pu. \end{array}$$

- *Repetition*: $f'(x)$ stands for the first derivative and $f''(x)$ denotes the second derivative. The strokes are reinterpreted as Roman numerals in $f^{(k)}(x)$, the derivative of order k.

3 Conventions for the use of mathematical symbols

Mathematical symbols are *conventions*. This can be seen best at the fact that one can use a different notation to express the same idea. The assertion

$$\frac{\mathrm{d} \sin x}{\mathrm{d}x} = \cos x$$

can be expressed equivalently as $\sin' y = \cos y$.

Although the freedom to use an arbitrary notation has no limits, conventions and rules are very important. There are good reasons for such behaviour which are important from an educational viewpoint. A steady change of notation impedes communication. A carefully chosen symbolisation may shed light on connections and reduce the labour of memory. There are some widely accepted notations.

- π for the circle number and e for the base of natural logarithms
- the use of lower case letters as variables for numbers
- the use of the Greek letters ε and δ for "small" numbers
- the use of symbols for relations like $=, <, >, \leq, \geq$
- the meaning of the algebraic symbols $+, -, \cdot, :, \sum, \prod$, of the root symbol $\sqrt{\ }$, and the logical symbols $\vee$, $\wedge$, $\neg$, $\Rightarrow$, $\Leftrightarrow$, $\exists$, $\forall$
- the use of the symbol $\parallel$ "parallel", $\perp$ "perpendicular", $\sim$ "similar", $\cong$ "congruent" (in geometry), $\equiv$ "identically equal", "congruent" (in algebra)

Such conventions are widely distributed. However, there are some rules which resemble the rules of the grammar of a language. What follows should give some ideas in the description of the "implicit" grammar of mathematical symbolism. The notion "implicit" means that these rules in most cases are not taught explicitly, but are followed like the rules of grammar.

3.1 SERIALISATION

To assist the memory it is useful to resort to ordered data. This can be the sequence of natural numbers or the sequence of signs of an alphabet. The order of some subsequences is old cultural heritage. The Hebrew alphabet starts with *aleph* א, *beth* ב, *gimel* ג and the Greek alphabet with α, β, γ. In the Arabic culture the older order of the alphabet also was *alif* ا, *bâ* ب, *ǧîm* ج. The order a, b, c reflects the fact that Latin c originally denoted a velar stop (close to k or g). The subsequence k, l, m, n has also survived some millennia.

The order of the various alphabets was fixed enough that these signs were also used for numbers. As late as 1617 J. Napier used the sequence $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, ... as a dyadic code, e. g. $1611 = 2^{10} + 2^9 + 2^6 + 2^3 + 2^1 + 2^0$ was represented as ***lkgdba*** (obviously Napier had $i = j$). The notation $\alpha = a + \mathbf{i}b + \mathbf{j}c + \mathbf{k}d$ for a quaternion clearly reflects this idea.

The letter x seems to be the most common device for an unknown number or a variable. If more variables are used one chooses the next letters y and z. If more letters are necessary very often one chooses a new subsequence e. g. u, v, w. Clearly another device is to use $x_1, x_2, x_3, \ldots$ As the last number of a count is the number of counted items, it would be a little be strange to use x_2, x_3, x_6 in a system of equations with three unknown quantities. One can also use a notation like $a_i, a_{i+1}, a_{i+2}, a_{i+3}, \ldots$ Clearly, the system may be disturbed by the fact that some letters have a connotation in the context. If the letter e is used for Euler's number then a sequence of constants a, b, c, d must stop here! The sign π very often is fixed by its meaning as the circle number. However, π, ρ, σ, ... are used for permutations in group theory. Note that this block can be found in exactly the same order in the Greek alphabet where $\pi = 80$, $\rho = 100$, $\sigma = 200$ (to represent the number 90 a special sign called *koppa* was used).

Viète used a quite different system. The letters for vowels were used for unknown quantities and the letters for consonants for known quantities. His famous rule for the connection between the coefficients of a quadratic equation and the roots was written as follows:

"Si $\overline{B+D}$ in A—A quad., aequaliter B in $D : A$ explicabilis est de qualibet illarum duarum B vel D." ('The equation $(B+D)A - A^2 = BD$ has the roots B and D'. Note the line over the symbols was used for the bracket and the Latin 'in' stands for multiplication).

Bhāskara used the words for colours (and their first letters) to denote unknown quantities extending the first one x_1 (which was called *yāvat tāvat*), namely *kālaka* 'black', *nīlaka* 'blue', *pīlaka* 'yellow' and *lohitaka* 'red'.

Serialisation helps to memorise but it also increases the readability of a text as a kind of "advanced organizer". If one finds in a text the notation V for a vector space and suddenly one reads W, in most cases this letter denotes another vector space. If a text uses the letters f and g for continuous functions, a further function will very often be denoted by h. However, in most cases one chooses ψ after ϕ, although in the Greek alphabet the next letter would be χ.

4 Configurations

There are some rules which generate "good" configurations. One rule may be called *similarity within a configuration.* A notation which mixes numeration like $x_1, x^2, \ldots$ in a sequence would be seen as strange. The same would apply to the use of x, Y, ζ instead of x, y, z. Clearly there are some exceptions. An example is the notation $s = \sigma + it$ for complex numbers in analytic number theory. In this case the rule of *alphabetic correspondence* has won. σ denotes the real part of s, similar to the notation $\alpha = a + ib$ and $\gamma = c + id$ where α corresponds to a and γ corresponds to c. Traditionally, the vertices of a triangle will be denoted by A, B, C, the opposite sides by a, b, c and the angles by α, β, γ. However, for a rectangle a different system has to be used!

Alphabetic correspondence is used in connection with diacritic signs. The derivative of a function f can be denoted as f' . Then Leibniz's rule $(fg)' = f'g + fg'$ is easy to remember. In a similar way the primitive function of f will be denoted as F. The dual space of a vector space V is denoted as $V*$.

But there is also a *rule of contrast.* When you use capital letters for points then probably you will choose lower case letters for lines. If you need a further notation for planes you could take the Greek alphabet. In the equation of a line $ax + bx + c = 0$ the variables x and y contrast with the other variables a, b, and c (in this context often called parameters). This rule of contrast is not followed in physics which makes the formulas less readable! A good example is the equation of planetary motion

$$\frac{\mathrm{d}r}{\mathrm{d}\varphi} = \frac{mr^2}{j}\sqrt{\frac{2}{m}(E + \frac{\gamma mM}{r}) - \frac{j^2}{m^2r^2}}.$$

Alphabetic correspondence can be seen in the notation m and M for the masses involved, r for the distance (derived from radius) and $E$for energy (j stands for angular momentum, γ for the gravitational constant). A similar case is VAN DER WAALS' equation $\left(p + \frac{a}{V^2}\right)(V - b) = RT$, where we find p for pressure, V for volume and T for temperature, and R a thermodynamic constant. A mathematician would like to see the equation $\left(x + \frac{a}{y^2}\right)(y - b) = Rz$!

Sometimes a conflict appears: If one denotes a point in the plane by $X = (x, y)$ then the principles of alphabetic correspondence and of serialisation can produce different continuations $X = (x_1, x_2)$, $Y = (y_1, y_2)$ or $X_1 = (x_1, y_1)$, $X_2 = (x_2, y_2)$ as notations for two points in the plane.

Alphabetic correspondence is also the source of new notations. The sum of two numbers is denoted by the symbol + and the product by a cross x or a dot $\cdot$ or very often suppressed at all, as in $2a$ or by an asterisk $*$. Note that multiplication by 1 is generally suppressed: We

write the letter a for $1a$. This is similar to the 1-deletion with number words. We say *ten* instead $*$*one ten* but *one million* for $*$*million*! For the sum of several summands one uses the sign $\sum$ (capital **s**igma as **s**um) and for the product of several factors the symbol Π (capital **p**i as **p**roduct). Acronymic devices are very old. In ancient Greek in one of the numeral systems the capital letters Π, Δ, H were used for the numbers 5 (= pente), 10 (= deka), and 100 (= hekaton). On the computer we find: F **f**ormat, H **h**elp, S **s**ave etc. As already mentioned the symbol ∂ is just a variant of the letter d. In complex variables the symbol $\wp$ (a hand written p) is used for the double **p**eriodic functions. Intersection and union of two sets are expressed by the use of $\cap$ and $\cup$. For an arbitrary family of sets we use the same symbols but modified to capital letters: $\bigcap$and $\bigcup$. In algebra the sign $\prod$ for product has been extended to the sign $\coprod$ denoting the coproduct.

Symmetry is a peculiar form of correspondence. This correspondence can be a kind of pairing: the image $z = x + iy$ will be denoted as $w = u + iv$. The partial derivative operators $\dfrac{\partial}{\partial x}\dfrac{\partial}{\partial y}, \ldots$ correspond to the differentials $\mathrm{d}x, \mathrm{d}y, \ldots$ Brackets are always used in pairs: $(\ldots)$, $[\ldots]$ und $\{\ldots\}$. A notation like $a(b-c$ or $a(b-c]$ would be look strange. Only the expression $a(b-c)$ would be called well formed. Brackets are not necessary in all cases as can be illustrated by the Polish notation $abc-*$ (brackets are then necessary to distinguish $(52)(33)6-*$ from $5(233)6-*$). The expression $y = F(x)$ is just convention but the notation $y = F(x$ would serve the same purpose (note the wording "f-of-x" does not reflect the closing bracket). Some people would prefer $\dfrac{x+2}{x^2+4}$ contrasting with the expression $\dfrac{x+2}{4+x^2}$.

4.1 WELLFORMEDNESS

The syntax of mathematical texts obeys some principles of *wellformedness*. We note three such rules: *congruence*, *closure*, and *position*. The equation

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

follows the rule of congruence which says that the variable k must appear at least twice. The expression

$$\sum_{k=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$$

does not obey this rule (or the formula is wrong).

A rule of position says that the symbol = appears between at least two expressions. The expression $a+b=c$ is correct but the expression $ab+c=$ is incorrect or at least incomplete. A rule of closure would demand that the expression $\int f(x)$ should be completed to the expression $\int f(x)\,\mathrm{d}x$.

Bound variables must not be used as free variables within the same expression. The writing $\int \sin x\,\mathrm{d}x = -\cos x$ is not seen as correct but is sometimes tolerated as an "abus de langage".

Since variables are like pronouns the same letter may be used in different expressions. The formulae $\int \sin x\,\mathrm{d}x = -\cos y + C$ and $\int \frac{1}{x}\mathrm{d}x = \ln y + C$ can appear in the same text although the letter x cannot have the same connotation in both expressions. In the second example the case $x = 0$ is excluded.

References

- Daniels, P. T, Bright, W. (eds.), 1996, *The World's Writing Systems*. New York–Oxford : Oxford University Press.
- Davis, G., Hunting, R. P., 1990, *Language Issues in Learning and Teaching Mathematics*. Bundoora : La Trobe University.
- Gelb, I. J., 1963, *A Study of Writing*. University of Chicago Press.
- Gericke, 1984, *Mathematik in Antike und Orient*. Berlin u. a. : Springer-Verlag.
- Haarmann, H., 1991, *Universalgeschichte der Schrift*. Frankfurt–New York : Campus.
- Halliday, M. A. K., 1974, "Some aspects of sociolinguistics" in *Nairobi Report. Interactions between Linguistics and Mathematical Education.* Final Report of the Symposium sponsored by UNESCO, CEDO, and ICMI. Nairobi/Kenya/September 1–11 UNESCO: ED – 74/CONF, 808.
- Menninger, K., 1979, *Zahlwort und Ziffer*. Göttingen : Vandenhoeck & Ruprecht.
- Maier, H., Schweiger, F., 1999, *Mathematik und Sprache* (Mathematik für Schule und Praxis Band 4. Hrsg. von H.-Ch. Reichel) Wien : öbv&hpt.
- Tropfke, J., 1980, *Geschichte der Elementarmathematik* 1. Bd. *Arithmetik und Algebra* (neu bearbeitet von K. Vogel, K. Reich und H. Gericke). Berlin ua. : De Gruyter