EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

Memorandum 1977-09.
Issued August 1977.

Notation for concatenation.

by

N.G. de Bruijn.

University of Technology
Department of Mathematics
P.O.Box 513, Eindhoven.
The Netherlands.

Notation for concatenation.

by N.G. de Bruijn.

1. Introduction. The purpose of this note is to develop a notational system for mathematical discussion of languages. We wish to avoid the confusion that comes from two sources: (i) if letters or words are considered as mathematical objects, we want to give names to them which are again letters or words, and (ii) it is tempting (but objectionable) to denote concatenation of two words by means of concatenation of names for these words ("if a and b are words, then ab is a word"). What these two sources of confusion have in common is that the language (i.e. the letters, words, etc. we study) get confused with the metalanguage (i.e. the mathematical symbols by means of which we discuss the language).

A popular system that uses (ii) is Backus' normal form for the definition of programming languages. It can be applied in simple cases, and as long as language and metalanguage do not get too entangled. It certainly fails if we want to treat language theory a bit more formally. (It fails already if we want to make a distinction between a letter and a one-letter word!) It is definitely inferior to the system that is used in the theory of context-free languages (cf. Ginsburg [1]). The latter system is certainly safe with respect to (i) and (ii). Nevertheless some trouble remains there: in order to narrow the gap between mathematics and interpretation, one easily gets into the habit of omitting the concatenation symbols, especially when discussing examples, and it is also in the examples that (i) reappears.

It is the purpose of this note to develop a style and a notation by means of which these kinds of confusion are avoided. In order to make things clear we shall postpone the use of terms like "word", "alphabet" until section 3, when we definitely know what we are talking about.

2. The sets A, $S_k(A)$ and $S(A)$. Let A be any set. $\mathbb{N}$ is the set of non-negative integers. For any $k \in \mathbb{N}$ the set $\mathbb{N}_k$ is the set of all $j \in \mathbb{N}$ with $j < k$ (hence $\mathbb{N}_0$ is empty). For every $k \in \mathbb{N}$ we denote by $S_k(A)$ the set of all mappings of $\mathbb{N}_k$ into A. In particular $S_0(A)$ contains just one element: the mapping that maps the empty set into A. Let us call that element ϵ, whence $S_0(A) = \{\epsilon\}$. And we define S by

$$S(A) = \cup_{k=0}^{\infty} S_k(A).$$

If $v \in S(A)$ then there is just one $k \in N$ with $v \in S_k(A)$; that k will be called the length of v.

There is a natural one-to-one mapping σ_1 of A onto $S_1(A)$: if $x \in A$ then $\sigma_1(x)$ is the mapping of N_1 into A that maps 0 onto x.

If $v \in S(A)$, $w \in S(A)$, and if v and w have length k and m, respectively, then we can form the mapping z of N_{k+m} into A, defined by

$$z(i) = v(i) \qquad (0 \leq i < k), \qquad z(i) = w(i-k) \qquad (k \leq i < k+m).$$

This z is called the concatenation of v and w. We denote it as follows:

$$z = \overline{v \mid w}. \tag{2.1}$$

Note that this definition leads to

$$\overline{v \mid \varepsilon} = \overline{\varepsilon \mid v} = v$$

for all $v \in S(A)$. And it is not hard to prove the associative rule: if $u,v,w \in S(A)$, and

$$\overline{u \mid v} = p, \quad \overline{v \mid w} = q \quad \text{then} \quad \overline{p \mid w} = \overline{u \mid q}.$$

We write $\overline{u \mid v \mid w}$ for that common value. Similarly we define $\overline{u_1 \mid u_2 \mid u_3 \mid u_4}$, etc. Let us call this the comb notation.

We also write

$$\overline{\prod_{i=1}^{n} f(i)} \tag{2.2}$$

if $n > 0$ and if f is a mapping of $\{1,\dots,n\}$ into S(A). What we intend is, of course, this: if (2.2) is denoted by k(n) then $k(1) = 1$ and $k(n) = \overline{k(n-1) \mid f(n)}$ $(n = 1,2,\dots)$. Note that $k(3) = \overline{f(1) \mid f(2) \mid f(3)}$, etc.

The case n=1 is a superfluous notation: $\overline{x} = x$ for all $x \in S(A)$. Nevertheless we may sometimes use it in cases like in (3.3).

We shall also use the notation

$$\overline{P \mid Q \mid r} = \{\overline{p \mid q \mid r} \mid p \in P, q \in Q\}$$

if $P \subset S(A)$, $Q \subset S(A)$, $r \in S(A)$, and the similar notation for other combinations of elements and subsets of S(A). In particular, $\overline{P} = P$ for all subsets P of S(A). And

$$\overline{\prod_{k=1}^{m} P_k} = \{\overline{\prod_{k=1}^{n} p_k} \mid p_1 \in P_1, \dots, p_n \in P_n\}$$

if $P_1,\ldots,P_n$ are subsets of S(A).

If f is a mapping of $\mathcal{P}(S(A))$ (i.e. the set of all subsets of S(A)) into $\mathcal{P}(S(A))$, and if f satisfies, for all sequences $P_1,P_2,\ldots$

$$f(P_1) \cup f(P_2) \cup \ldots = f(P_1 \cup P_2 \cup \ldots) \tag{2.3}$$

then the set

$$f(\emptyset) \cup f(f(\emptyset)) \cup f(f(f(\emptyset))) \cup \ldots$$

is the minimal solution of the equation

$$X = f(X) \qquad (X \in \mathcal{P}(S(A))). \tag{2.4}$$

Note that (2.3) implies monotonicity: if $P \subset Q$ then $f(P) \subset f(Q)$.

If f satisfies (2.3), if X is the minimal solution of (2.4), and if we want to prove some proposition for all $x \in X$, it suffices to show the following for all $Y \in \mathcal{P}(S(A))$: if the proposition holds for all $x \in Y$ then it holds for all $x \in f(Y)$. Such a proof is called a "proof by recursion" with respect to (2.4). A similar arrangement can be made for "definitions by recursion".

What has been said on (2.4) can be generalized to the case of several variables, where we have a set of equations like $X = f_1(X,Y), Y = f_2(X,Y)$.

We consider a special case (of 2.4) that occurs quite often. Let $P \subset S(A)$, then the minimal solution of

$$X = P \cup [X\,P]$$

is $P \cup [P\,P] \cup [P\,P\,P] \cup \ldots$. We denote it by str(P). If $q \subset S(A)$, we use str(P,q) to denote the minimal solution of

$$X = P \cup [X\,q\,P].$$

Hence $str(P,q) = P \cup [P\,q\,P] \cup [P\,q\,P\,q\,P] \cup \ldots$.

3. Application to the description of languages. We want to discuss formal languages by means of the framework presented in section 2. Let us call the basic set A the alphabet, and its elements letters or symbols. The elements of S(A) are called words; in particular ε is called the empty word, and the elements of $S_1(A)$ are called one-letter words or atoms. Yet we keep using letters, or combinations of letters and other symbols, for denoting (in the metalanguage) elements of A or S(A), according to standard traditions in mathematics. Let us agree not to try to write the letters of the alphabet A themselves; we just stick to their denotations in the metalanguage. And, what helps to reduce danger

of confusion, we usually prefer to talk about elements of S(A) rather than about elements of A (the reader may notice that in section 2 the elements of A are hardly mentioned).

If we want to say that we have a two-letter alphabet, we cannot just say "let A = {a,b}" since this does not exclude the possibility a = b. We have to say: "let A = {a,b}, where $a \neq b$". Having said this,the symbols a and b have the same distinctive power as "if we had the letters themselves". The situation is not very different from what we do when introducing a set of two points in a plane. Our mathematical discussion will contain the names, and not "the points themselves". If we say "let x be a letter" then this means just the same thing as "let x be an element of A" and has nothing to do with the fact that in another sense x is a letter already.

We get a sound mathematical notational system this way, but apart from soundness we require a system to have other features, like:(i) it should be suggestive, and (ii) it should not be tedious.

Quite often we want to think of some elements of the alphabet being symbols other than letters, and in those cases we dislike denoting them by letters or words. In current mathematical notation the only names we use for variables and constants are letters or words (possibly with numbers as indices). It is, of course, possibly to say a thing like this: "the alphabet has six elements, viz. a,b,c (called letters), op (called opening parenthesis), cp (called closing parenthesis) and co (called comma)", but is is awkward to talk about

$$\boxed{\rho(a)}\ \boxed{\rho(op)}\ \boxed{\rho(b)}\ \boxed{\rho(co)}\ \boxed{\rho(c)}\ \boxed{\rho(cp)} \tag{3.1}$$

if what we really have in mind is a(b,c). It is a bit better already if we agree not to use a,b,c,op,cp,co for the elements of A but for the corresponding one-letter-words (i.e. the elements of $S_1(A)$); then we get

$$\boxed{a}\ \boxed{op}\ \boxed{b}\ \boxed{co}\ \boxed{c}\ \boxed{cp}\ . \tag{3.2}$$

We can improve on this. Bearing in mind that $\boxed{x} = x$ for all $x \in S(A)$, we can take care that the name "op" never appears outside a comb, and if it occurs under a comb it is all by itself in a compartment. This is why we can allow an exception to the rule that variables and constants have to be denoted by letters or words. It does not do any harm to write: "let A have six elements, and let the atoms be

$$a, \quad b, \quad c, \quad \boxed{(}, \quad \boxed{)} \quad \text{and} \quad \boxed{,}\ ". \tag{3.3}$$

Then we can write, instead of (3.1) and (3.2)

$$\boxed{a\,|\,(\,|\,b\,|\,,\,|\,c\,|\,)\,|} \quad .$$

Quite often we have long combed formulas to discuss, and we of course would like to get rid of the combs. This can often be achieved by the following convention. If we write a thing like

$$\sin(f(\cos(x))) \tag{3.4}$$

and we proclaim it to be uncombed, then what we mean is

$$\boxed{\sin\,|\,(\,|\,f\,|\,(\,|\,\cos\,|\,(\,|\,x\,|\,)\,|\,)\,|\,)\,|} \quad . \tag{3.5}$$

We get from (3.4) to (3.5) if we consider the maximal subsequences of consecutive letters in (3.4) as names of elements of S(A).

The uncombed version is not always possible. For example, in the case of the six atoms (3.3) we may write: "Let f be a mapping of {a,b,c} into {b,c}; we consider the word w = $\boxed{a\,|\,(\,|\,f(a)\,|\,)}$, which is either $\boxed{a\,|\,(\,|\,b\,|\,)}$ or $\boxed{a\,|\,(\,|\,c\,|\,)}$ ". Here we cannot use an uncombed form for w, since there is one compartment that contains both letters and other symbols. The notation "a(f(a)) (uncombed)" would mean $\boxed{a\,|\,(\,|\,f\,|\,(\,|\,a\,|\,)\,|\,)\,|}$.

We might extend the possibilities of uncombed presentation if we use different kinds of parentheses for different purposes, but then the disadvantages of the combs are replaced by other notational intricacies. We do not discuss this kind of thing here. We neither look into the situation where empty spaces are considered as part of the language, and we do not ask what it means to denote empty spaces by empty spaces in the metalanguage.

Some of the trouble we have when trying to make our notation suggestive, is that standard mathematical notation is not always consistent. E.g.: sin might be a single identifier but also a notation for a product of three numbers.

4. An example. (We want to describe formulas like f(g(x,y),h(u),c)).

Let us agree that A is such that the set $S_1(A)$ of all one-letter words contains the following five mutually disjoint subsets:

$$C, \qquad V, \qquad \{\boxed{(}\}, \qquad \{\boxed{)}\}, \qquad \{\boxed{,}\}.$$

(the latter three have one element each). We now define the subset F of S(A) as the minimal solution of

$$F = C \cup V \cup \boxed{C\,|\,(\,|\,\mathrm{str}(F,\boxed{,}\,|\,)\,|\,)\,|} \quad . \tag{4.1}$$

Let us assume that C contains f,g,sin,c and that V contains x,y. Then some of the things that F contains are (uncombed)

g(sin(f(u,g(c)))

g(x,y,u,f(sin(x))

sin(f(sin,x),y,g(x,y))

f(sin(f,y),y,g(y,y))

These words have length 14,15,18,18 respectively. Note that we have not required f,g,sin,c to be different, nor that $x \neq y$. So it makes sense to say that if f=sin and x=y then the last two words are equal.

5. A remark on parsing. The set F of section 4 has the property that for every $w \in F$ there is exactly one way to build it up from smaller words of F : w is either in C, or in V or it has the form $\lceil p \mid (\mid s \mid) \rceil$ with $p \in C$, $s \in \mathrm{str}(F, \lceil , \rceil)$, with p and s uniquely determined. And if $s \in \mathrm{str}(F, \lceil , \rceil)$ then either $s \in F$ or s has the form $\lceil s_1 \mid , \mid q \rceil$ with $s_1 \in \mathrm{str}(F, \lceil , \rceil), q \in F$. These unique parsing properties would still hold if we suppress the commas, i.e. if in (4.1) $\mathrm{str}(F, \lceil , \rceil)$ is replaced by str(F). An example would be

$$\lceil g \mid (\mid \sin \mid (\mid f \mid (\mid u \mid g \mid (\mid v \mid) \mid) \mid) \rceil ,$$

but we would not be able to present it in an uncombed form since ug would be read as a single name!

Reference.

1. S. Ginsburg. The mathematical theory of context-free languages. McGrawHill 1966.