On a theorem of Hanna Neumann

By SAVA KRSTIĆ (Belgrade)

Let $W(x, y)$ be a group word in x, y and element of a free group G. HANNA NEUMANN proved in [3] that such a word determines an associative operation on G if and only if $W(x, y)=a, x, y, xay$ or yax, for some constant a from G. In [2] one can find a generalization of this theorem. Our goal in the present article is to generalize H. Neumann's theorem in another direction. Namely, we shall consider the *equation of generalized associativity*

$$A(x, B(y, z)) = C(D(x, y), z) \tag{1}$$

and find among all quadruples (A, B, C, D) of group words in two variables over a free group those for which the above equation holds.

Refering the reader to the book [1] we only remark that functional equations of associativity and generalized associativity have been solved in many different contexts.

Finally, we wish to thank A. KRAPEŽ, for the present article originated in a discussion with him.

§ 1. Preliminaries

Polynomials. In accordance with the corresponding concept of universal algebra, by a (group) *polynomial* we mean any group word $P(x_1, \ldots, x_n)$ built of variables $x_1, \ldots, x_n$ and elements of the group G. Clearly, the just defined polynomials over G are elements of the free product of G and the free group $\langle x_1, x_2, \ldots\rangle$. Every polynomial $P(x_1, \ldots, x_n)$ determines an n-ary *polynomial operation* on G. Without many efforts one can prove that in the case of a free group G the correspondence between polynomials and polynomial operations is one-to-one. Since we shall work under restriction to that case we shall use in the sequel only the word polynomial giving it both mentioned meanings.

As usual, we say that the polynomial $P(x_1, \ldots, x_n)$ is *non-degenerate* if it really depends on all variables $x_1, \ldots, x_n$.

Every polynomial can be represented as a product of monomials:

$$P(x_1, \ldots, x_n) = P_1(\xi_1) \ldots P_k(\xi_k), \tag{2}$$

$\xi_1, \ldots, \xi_k$ belonging to the set of the x_i's. This representation is to be called a *factorization* of P if all monomials $P_1, \ldots, P_k$ are non-degenerate and if $\xi_i \neq \xi_{i+1}$ for every i, $1 \leqq i < k$. Polynomials need not have unique factorization, but if $P(x_1, \ldots, x_n)=Q_1(\eta_1) \ldots Q_{k'}(\eta_{k'})$ is another factorization of the polynomial P from

(2) then $k=k'$ and $\xi_i=\eta_i$, $1\leqq i\leqq k$ must hold. Thus the word $\tau P=\xi_1 \dots \xi_k$ is independent of the factorization chosen and it will be called the *type* of the polynomial P.

Subsidedness. Let

$$P(x_1, \dots, x_n) = a_0\xi_1^{\varepsilon_1}a_1\dots\xi_k^{\varepsilon_k}a_k \tag{3}$$

be in reduced form as an element of $G*\langle x_1, \dots, x_n\rangle$ The *subsidence* P° of the polynomial P is, by definition, the polynomial obtained from the right-hand side of (3) by deleting the symbols a_0, a_k and all those a_i for which $\xi_i\neq\xi_{i+1}$. The polynomial P is *subsided* if $P=P^\circ$. Clearly, P is subsided iff it factorizes as a product of subsided monomials; subsided polynomials have unique factorization. Further, if (2) is a factorization of P then

$$P^\circ(x_1, \dots, x_n) = P_1^\circ(\xi_1)\dots P_k^\circ(\xi_k)$$

is the factorization of P°. Also, subsidedness is preserved under substitution and multiplication of polynomials.

Later on we shall make use of the following simple

Lemma. *Let U, V, W be subsided non-degenerate monomials over a free group G. If $\tau U(V(x)W(y))$ is equal to xy, yx, xyx, yxy, then $U(t)$ is equal respectively to $t, t^{-1}, tpt^{-1}, t^{-1}pt$, where p is an element of G.*

Similar solutions. For a given $U(x, y)$ denote by $\overline{U}(x, y)$, $U'(x, y)$ and $U''(x, y)$ the polynomials $U(y, x)$, $U(x^{-1}, y)$ and $U(x, y^{-1})$. Given a solution (A, B, C, D) of the equation (1) one can immediately obtain the three related solutions $(\overline{C}, \overline{D}, \overline{A}, \overline{B})$, (A'', B^{-1}, C, D) and (A, B, C', D^{-1}) called respectively *dual, left transform* and *right transform* of (A, B, C, D). Two solutions of (1) are to be called *similar* if one can pass from one of them to another applying several times the operations of dualizing and left and right transforming. Similarity is evidently an equivalence relation and it is easy to prove that there are at most eight different solutions equivalent to a given one. (In fact, the group of similitudes which acts on the set of all solutions of (1) is isomorphic to the group of quaternions.)

§ 2. The Theorem

There is a lot of degenerate solutions of the functional equation (1). Since it is both simple and a cumbersome task to find and write down all of them we shall confine our attention to the non-degenerate case; so, the polynomials A, B, C, D are supposed to depend actually on two variables.

Further, because a solution of the equation (1) immediately produces all eight that are similar to it, we shall work "up to similarity". The list of solutions is considered complete if it contains at least one member of every class of similar solutions.

Finally, we shall work under one more restriction which is of somewhat greater complexity. Note that if (A, B, C, D) is a solution of (1) then its subsidence $(A^\circ, B^\circ, C^\circ, D^\circ)$ is a solution too. There is a close relationship between two solutions with the same subsidence because they differ only in several constants out of G. Thus the essential part of a solution is contained in its subsidence. To avoid

introducing a number of parameters (which would be indispensable if one wished to have all solutions) we are going to find only subsided solutions. We remark that the method used for obtaining subsided solutions does not need any substantial refinement in order to handle the general case. As an example (see Corollary 2) we shall write out the solutions with a complete set of parameters in a simpler case of solving equation (1) over a free semigroup.

Theorem. *Every subsided non-degenerate solution of the equation* $A(x, B(y, z)) = C(D(x, y), z)$ *over a free group* G *is similar to a solution of one of the following two forms:*

(I)
$$A(x, t) = (Pt)^{\alpha_0} p_1 (Pt)^{\alpha_1} p_2 \dots p_n (Pt)^{\alpha_n}$$
$$B(y, z) = QR$$
$$C(s, z) = (sR)^{\alpha_0} p_1 (sR)^{\alpha_1} p_2 \dots p_n (sR)^{\alpha_n}$$
$$D(x, y) = PQ$$

(II)
$$A(x, t) = Pt^{\alpha_0} P^{-1} p_1 Pt^{\alpha_1} P^{-1} p_2 \dots p_n Pt^{\alpha_n} P^{-1}$$
$$B(y, z) = QRQ^{-1}$$
$$C(s, z) = sz^{\alpha_0} s^{-1} p_1 sz^{\alpha_1} s^{-1} p_2 \dots p_n sz^{\alpha_n} s^{-1}$$
$$D(x, y) = PQ,$$

where $P = P(x)$, $Q = Q(y)$, $R = R(z)$ *are subsided monomials,* $n \geqq 0$, $p_1, \dots, p_n$ *are non-trivial elements of* G *and* $\alpha_0, \dots, \alpha_n$ *a sequence of non-zero integers vhich in the case* (I) *satisfies the additional condition that any two consecutive members of it are of different sign.*

PROOF. Let $A(x, t)$, $B(y, z)$, $C(s, z)$, $D(x, y)$ be subsided monomials and let

$$T(x, y, z) = A(x, B(y, z)) = C(D(x, y), z),$$

the ternary polynomial T is then subsided too.

We begin with deducing some information about types of polynomials occurring in the equation above.

(4) τT does not begin with y.

Assume the contrary; then $\tau A(x, t)$ begins with t and the first factor in the factorization of $A(x, t)$ is a monomial $U(t)$. $\tau U(B(y, z))$ is an initial segment of τT which implies that in τT both y and z occur before the first occurence of x. Analogously we obtain that $\tau C(s, z)$ begins with s and thence that both x and y occur in τT before the first occurence of z. The contradiction just obtained establishes (4).

By a similar argument one can prove that

(5) None of xyx and zyz occurs as a subword of τT.

If $U(t)$ is a non-degenerate monomial then $\tau U(V(x_1, \dots, x_k))$ contains as a subword at least one of $\tau V(x_1, \dots, x_k)$ and $\tau V^{-1}(x_1, \dots, x_k)$. This general fact and (5) yield that zyz does not occur as a subword in τB and that xyx does not

occur in τD. Thus we get $\tau B=yz$, zy or yzy and $\tau D=xy$, yx or yxy. The proof of (4) gives also that τT cannot begin with xz or zx which, together with (5), implies that τT begins with xyz or zyx. Hence the two equalities $\tau B=yzy$ and $\tau D=yxy$ cannot hold at the same time.

Utilizing duality one can pass from a solution in which $\tau D=yxy$ holds to a solution with $\tau B=yxy$. Further, the left (right) transform of a solution in which $\tau D=yx$ ($\tau B=zy$) holds is a solution with $\tau D=xy$ ($\tau B=yz$). So, as for the types τB and τD it is enough up to similarity to consider only two cases.

Case 1 $\tau B=yz$ and $\tau D=xy$.

Let $B(y, z)=Q_1(y)R(z)$ and $D(x, y)=P(x)Q_2(y)$ be factorizations for B and D. Since we do not know anything in advance about the types of A and C we write the factorizations

$$(6)\qquad \begin{cases} A(x, t) = P_1(x)A_1(t)\dots P_k(x)A_k(t) \\ C(s, z) = C_1(s)R_1(z)\dots C_{k'}(s)R_{k'}(z) \end{cases}$$

allowing some of the monomials P_1, A_k, C_1, R_k, to be trivial.

Our first goal is to prove $Q_1=Q_2$. Suppose $P_1\neq 1$. It follows that τT begins with x and then that it begins with xyz. Hence $\tau C_1\big(D(x, y)\big)=xy$ and by means of the Lemma we get $C_1(s)=s$. Similarly $\tau A_1\big(B(y, z)\big)=yz$ or yzy and the Lemma implies $A_1(t)=t$ or tpt^{-1}. In any case we have

$$A\big(x, B(y, z)\big) = P_1(x)Q_1(y)R(z)\dots$$

$$C\big(D(x, y), z\big) = P(x)Q_2(y)R_1(z)\dots.$$

All polynomials in these equalities are subsided and $Q_1=Q_2$ follows. The remaining case $P_1=1$ is quite analogous and leads to the same conclusion.

So we have $B(y, z)=Q(y)R(z)$ and $D(x, y)=P(x)Q(y)$ for some monomials P, Q, R.

From (5) it follows that between any two successive occurrences of x in $\tau T(x, y, z)$ there stands one of the words yz, zy, yzy. Invoking the Lemma we obtain that all monomials $A_i(t)$ (except perhaps $A_k=1$) in the factorization (6) are equal to t, t^{-1} or $t^{-1}p_it$, where the p_i are elements of G. Analogously, every factor $C_i(s)$ is equal to s, s^{-1} or $sp_i's^{-1}$. Consequently in the factorization of $T(x, y, z)$ every factor in x is of the form P, P^{-1} or $P^{-1}pP$, factors in y are Q or Q^{-1} and factors in z are R, R^{-1} or RpR^{-1}.

(7) Looking at $T(x, y, z)$ as $A(x, B(y, z))$ we see that immediately after any occurence of the factor Q in T stands an R or RpR^{-1}, immediately before any occurence of Q^{-1} stands an R^{-1} or RpR^{-1} and every RpR^{-1} is surrounded by Q and Q^{-1} following the pattern $QRpR^{-1}Q^{-1}$.

In the same manner it follows from $T(x, y, z)=C\big(D(x, y), z\big)$ that

(8) immediately after any occurence of the factor Q^{-1} in T stands a P or $P^{-1}pP$, immediately before any occurence of Q stands a P or $P^{-1}pP$ and every factor $P^{-1}pP$ occurs as a part of some $Q^{-1}P^{-1}pPQ$.

One can easily prove (see (4) and (5)) that neither xzx nor zxz can be a subword of τT. This implies that

(9) $\begin{cases} \text{among any three consecutive factors in the factorization of } T(x, y, z) \text{ at least} \\ \text{one is } Q \text{ or } Q^{-1}. \end{cases}$

Taking into consideration that τT begins and ends with xyx or zyx it follows from (7), (8) and (9) that the polynomial $T(x, y, z)$ is of the form

$$T(x, y, z) = (PQR)^{\alpha_0} p_1 (PQR)^{\alpha_1} p_2 \dots p_n (PQR)^{\alpha_n}, \tag{10}$$

where $p_1, \dots, p_n$ are nontrivial elements of G and $\alpha_0, \dots, \alpha_n$ are nonzero integers such that $\alpha_i \alpha_{i+1} < 0$, $1 \leqq i < n$.

Now the equality (10) directly implies that the polynomials A, B, C, D are exactly of the form described in part (I) of the Theorem.

Case 2. $\tau B = yzy$ and $\tau D = xy$.

We do not wish to repeat arguments from the previous case, so we only remark that now there exist monomials P, Q, R such that

$$B(y, z) = Q(y) R(z) Q^{-1}(y), \quad D(x, y) = P(x) Q(y),$$

and then one can deduce the equality

$$T(x, y, z) = PQR^{\alpha_0} Q^{-1} P^{-1} p_1 PQR^{\alpha_1} Q^{-1} P^{-1} p_2 \dots p_n PQR^{\alpha_n} Q^{-1} P^{-1}$$

from which the solution described in part (II) of the Theorem is obtained.

Corollary 1. *(Nontrivial part of H. Neumann's Theorem.) The nondegenerate solutions of the functional equation* $W(x, W(y, z)) = W(W(x, y), z)$ *over the free group* G *are* $W(x, y) = xay$ *and* $W(x, y) = yax$, *where* a *is an arbitrary element of* G.

Proof. If the quadruple (W, W, W, W) is a solution of the equation (1) then $(W^\circ, W^\circ, W^\circ, W^\circ)$ or its dual $(\overline{W}^\circ, \overline{W}^\circ, \overline{W}^\circ, \overline{W}^\circ)$ is of the form (I), since in every solution of the form (II), $B \neq D$. Now if we impose in (I) the condition $A = B = C = D (= U)$ we get immediately $U(x, y) = xy$.

If $W^\circ = U$ then $W(x, y) = bxayc$ for some constants a, b, c out of G. Putting this into the equation of associativity gives $b = c = 1$. The case $\overline{W}^\circ = U$ leads in the same way to the conclusion $W(x, y) = yax$.

Corollary 2. *Every non-degenerate solution of the equation* $A(x, B(y, z)) = C(D(x, y), z)$ *over a free semigroup* S *is equal or dual to a solution of the form*

$$\begin{aligned} A(x, t) &= c\, d'\, Pata_1\, Pata_2 \dots a_n\, Pata_{n+1} \\ B(y, z) &= b' Q\, d'' c R b'' \\ C(s, z) &= c_0 sc R c_1 sc R c_2 \dots c_n sc R b'' a_{n+1} \\ D(x, y) &= d' P a b' Q\, d'', \end{aligned}$$

where $P = P(x)$, $Q = Q(y)$, $R = R(z)$ *are non-degenerate monomials,* $n \geqq 1$ *and* $a, b', b'', c, d', d'', a_i, c_i$ $(0 \leqq i \leqq n+1)$ *elements of* S *such that* $b'' a_i = c_i d'$, $1 \leqq i \leqq n$.

PROOF. Of course, the solution of equation (1) in the semigroup case cannot be directly deduced from the Theorem. But it suffices to apply a simplified variant of the method used in proving the Theorem. Accordingly, the first thing is to solve (1) for subsided polynomials. Thus we obtain $T(x, y, z)=(PQR)^n$ or $T(x, y, z)=(RQP)^n$, $n\geqq 1$. The general solution is then obtained by introducing constants between monomials in factorizations and finding a dependence among them which is necessary and sufficient in order to provide fulfillment of the equation (1).

References

[1] J. ACZÉL, Lectures on functional equations and their applications. *New York, London: Academic Press*, 1966.

[2] S. KRSTIĆ, Two results on associativity of composite operations in groups. *Publ. Inst. Math. Beograd* **33** (47) (1982).

[3] H. NEUMANN, On a question of Kertész. *Publ. Math. Debrecen* **8** (1961), 75—78.

SAVA KRSTIĆ
MATEMATIČKI INSTITUT
KNEZ MIHAILOVA 35
11000 BEOGRAD
YUGOSLAVIA

(Received August 9, 1982.)