RECURSION DEPTH ANALYSIS FOR SPECIAL TREE TRAVERSAL ALGORITHMS

Peter Kirschenhofer, Helmut Prodinger
Institut für Algebra und Diskrete Mathematik
TU Vienna, A-1040 Vienna, Gußhausstraße 27-29, Austria

1. INTRODUCTION AND MAIN RESULTS

In this paper we are concerned with the analysis of special recursive algorithms for traversing the nodes of a planted plane tree (ordered tree; planar tree). Some by now classical results in this area are due to KNUTH [7], DE BRUIJN, KNUTH and RICE [1], FLAJOLET [2], FLAJOLET and ODLYZKO [3], FLAJOLET, RAOULT and VUILLEMIN [4], KEMP [6] and others and are summarized in the next few lines:

The most important tree structure in Computer Science are the binary trees. The inorder traversal (KNUTH [7]) is the following recursive principle:

Traverse the left subtree
Visit the root
Traverse the right subtree.

The most straightforward implementation uses an auxiliary stack to keep necessary nodes of the tree. The analysis of the expected time of the visit procedure is clearly linear in the size of the input tree. To evaluate recursion depth means to determine the average stack height as a function of the size of the tree. The recursion depth or height h of the binary tree is recursively determined as follows: If the family $\mathcal{B}$ of binary trees is given by the symbolic equation

$$\mathcal{B} = \square + \underset{\mathcal{B}\quad\mathcal{B}}{\overset{\circ}{\wedge}}$$

then $h(\square) = 0$ and $h\Big(\underset{t_1\ \ t_2}{\overset{\circ}{\wedge}}\Big) = 1 + \max\{h(t_1), h(t_2)\}$.

In [3] FLAJOLET and ODLYZKO determined the average value h_n of h in the family $\mathcal{B}_n$ of binary trees of size n to be

$$h_n \sim 2\sqrt{\pi n}.$$

The recursive visit procedure can be optimized in the case of binary trees by eliminating endrecursion: the resulting iterative algorithm keeps at each stage a list of right subtrees that still remain to be explored. The storage complexity of this optimized algorithm is easily seen to correspond exactly to the so-called left-sided height h^* defined by

$$h^*(\square) = 0,\ h^*\Big(\underset{t_1\ \ t_2}{\overset{\circ}{\wedge}}\Big) = \max\{1+h^*(t_1), h^*(t_2)\}.$$

Recalling that the rotation correspondence (KNUTH [7;2.3.2]) transforms a binary tree of n-1 internal (binary) nodes into a planted plane tree with n nodes, the average storage complexity of the optimized algorithm follows immediately by a result of DE BRUIJN, KNUTH and RICE [1] about the average height of planted plane trees:

$$h_n^* \sim \sqrt{\pi n},$$

where the index n again refers to the size of the trees.

It was already proposed in KNUTH's book to consider this kind of questions for other families of trees. Dealing with the family P of planted plane trees defined by

$$P = o + \underset{P}{\overset{\circ}{|}} + \underset{P\ P}{\overset{\circ}{\wedge}} + \underset{PPP}{\overset{\circ}{\wedge}} + \dots$$

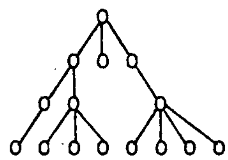
there are several meaningful analogues of the left-sided height of binary trees:

$$u(o) = 0;\ u\Big(\underset{t}{\overset{\circ}{|}}\Big) = u(t);\ u\Big(\underset{t_1 \dots t_r}{\overset{\circ}{\wedge}}\Big) = \max\{1+u(t_1),\dots,1+u(t_{r-1}),u(t_r)\},\ r\geq 2$$

$$v(o) = 0;\ v\Big(\underset{t_1 \dots t_r}{\overset{\circ}{\wedge}}\Big) = \max\{r-i+v(t_i)\,|\,1\leq i\leq r\}$$

$$w(o) = 0;\ w\Big(\underset{t}{\overset{\circ}{|}}\Big) = 1 + w(t);\ w\Big(\underset{t_1 \dots t_r}{\overset{\circ}{\wedge}}\Big) = \max\{1+w(t_1),w(t_2),\dots,w(t_r)\},\ r\geq 2$$

The "heights" u, w are better understood as follows. u counts the maximal number of edges not being a rightmost successor of a node in a chain connecting the root with a leaf. w counts the maximal number of edges which are leftmost successors of a node in a chain connecting the root with a leaf. For example we have for the tree t depicted below the values $u(t)=2$, $v(t)=4$, $w(t)=3$.



A short reflection tells us that u determines the recursion depth of the optimized tree traversal algorithm. (The non-optimized algorithm corresponds to the treatment of DE BRUIJN, KNUTH and RICE [1].)

The interpretation of v is a bit more complex: Recall that a binary tree can be used to represent arithmetic expressions; a simple strategy of the evaluation is "from right to left", i.e. to evaluate $\underset{t_1\quad t_2}{\overset{op}{\wedge}}$ we evaluate t_2, use one register to keep that value, evaluate t_1 and then perform "op". It is clear that the maximal number of registers during the evaluation of a binary tree is exactly h^*.

Planted plane trees are well suited to encode arithmetic expressions where k-ary operations may occur for any k. The same strategy as in the case of binary trees leads to v (evaluating $\underset{t_1 \dots t_r}{\overset{\circ}{\wedge}}$ from right to left during the consideration of t_i already $r-i$ registers are used to keep intermediate values).

The interest in w originates from another source; however since this parameter fits well in the concept of asymmetric heights we have decided to include it into our discussion.

In any instance we are interested in the average value u_n, v_n, w_n of the "height" u, v, w of the trees in P_n, i.e. the trees of size n in P. Our main results are:

THEOREM. a) $u_n = \frac{1}{2}\sqrt{\pi n} - 1 + O(n^{-1/2})$

b) $v_n = \sqrt{\pi n} - \frac{5}{2} + O(n^{-1/2})$

c) $w_n = \frac{1}{2}\sqrt{\pi n} + O(n^{1/4+\varepsilon})$, for all $\varepsilon>0$.

These results are achieved by means of a detailed singularity analysis of corresponding generating functions in the following section.

2. PROOFS AND MINOR RESULTS

Let $P_h(z)$, $U_h(z)$, $V_h(z)$, $W_h(z)$ be the generating functions of trees in P with ordinary height or "height" u, v, w, respectively, $\le h$ and $y(z)=(1-\sqrt{1-4z})/2$ the generating function of all trees in P. Then the generating functions of the sums of "heights" of trees of equal size are given by

$$\sum_{h\ge 0}(y-P_h),\ \sum_{h\ge 0}(y-U_h),\ \sum_{h\ge 0}(y-V_h) \text{ and } \sum_{h\ge 0}(y-W_h). \tag{1}$$

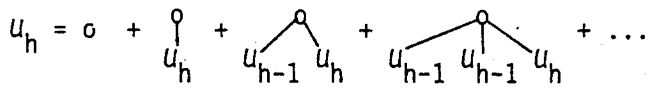
It is well known [1] that

$$P_0(z) = z;\ P_h(z) = z/(1-P_{h-1}(z)) \quad \text{and} \tag{2}$$

$$P_h(z) = \frac{u}{1+u}\cdot\frac{1-u^{h+1}}{1-u^{h+2}} \quad \text{where} \quad z = \frac{u}{(1+u)^2}. \tag{3}$$

LEMMA 1. $U_h(z) = P_{2h+1}(z)$.

Proof. We have $U_0(z) = \frac{z}{1-z}$ and because of



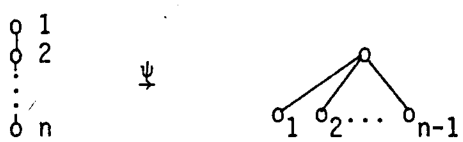
(with an obvious notation)

$$U_h(z) = z + \frac{zU_h(z)}{1-U_{h-1}(z)}, \text{ so that } U_h = \frac{z}{1-\frac{z}{1-U_{h-1}}},$$

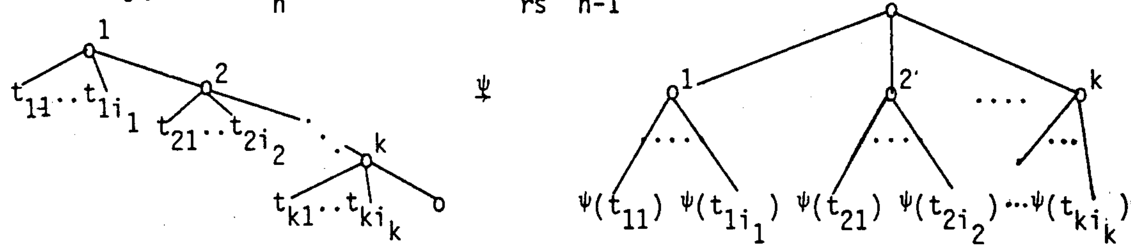
from which Lemma 1 follows immediately from (2) by induction.

An alternative proof can be given by defining the following map $\psi: U_h \to P_{2h+1}$ which turns out to be a bijection:

$\psi: U_0 \to P_1$ is defined by



and, recursively, for $t\in U_h$ with subtrees $t_{rs}\in U_{h-1}$



LEMMA 2. With $\mu(z) = 1-4z$ and some constants K_1, K_2 we have for $z \to 1/4$

$$\sum_{h\geq 0}(y-U_h) = K_1 - \frac{1}{8}\log\mu + \frac{1}{2}\mu^{1/2} + K_2\mu + \ldots$$

Proof.

$$\sum_{h\geq 0}(y-U_h) = \sum_{h\geq 0}(y-P_{2h+1}) = -\frac{u}{1+u} + \frac{1-u}{1+u}\sum_{h\geq 0}\frac{u^{2h+1}}{1-u^{2h+1}}$$

Now

$$\sum_{h\geq 0}\frac{u^{2h+1}}{1-u^{2h+1}} = \sum_{k\geq 1} d_1(k)u^k,$$

with $d(k) = d_1(k) + d_2(k)$, $d_2(2k) = d(k)$ where $d(k)$, $d_1(k)$, $d_2(k)$ denotes the number of all, odd or even divisors of k. So we have

$$\sum_{h\geq 0}(y-U_h) = -\frac{u}{1+u} + \frac{1-u}{1+u}\sum_{k\geq 0} d(k)u^k - \frac{1-u^2}{(1+u)^2}\sum_{k\geq 1} d(k)u^{2k}.$$

Now it is known [9] that

$$g(z) = \frac{1-u}{1+u}\sum_{k\geq 1} d(k)u^k = K_1' - \frac{1}{4}\log\mu + \frac{1}{4}\mu^{1/2} + K_2'\mu + \ldots \tag{4}$$

Since $u^2 = (\frac{z}{1-2z})^2 = 4\mu + O(\mu^2)$ it follows that

$$\frac{1-u^2}{(1+u)^2}\sum_{k\geq 1} d(k)u^{2k} = K_1'' - \frac{1}{8}\log\mu + \frac{1}{4}\mu^{1/2} + K_2''\mu + \ldots$$

Further $\frac{u}{1+u} = y(z) = (1-\mu^{1/2})/2$. Putting everything together the lemma follows. □

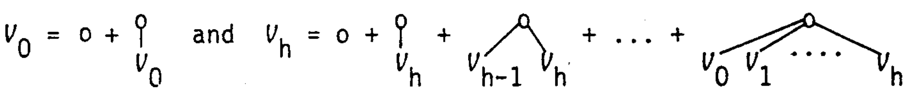
By a complex contour integration (compare [3]) the local expansion of Lemma 2 "translates" into the following asymptotic behaviour of the coefficients.

LEMMA 3. $\sum_{h\geq 0}(y-U_h) = \sum_{n\geq 0} z^n 4^n\left(\frac{1}{8n} - \frac{1}{4\sqrt{\pi}}\frac{1}{n^{3/2}} + O(\frac{1}{n^2})\right)$.

Dividing by $|P_n| = \frac{1}{n}\binom{2n-2}{n-1} = \frac{1}{4\sqrt{\pi}}4^n n^{-3/2}(1 + O(\frac{1}{n}))$ we achieve part a) of our main theorem.

LEMMA 4. $V_h(z) = P_{h+1}(z)$.

Proof. In the same style as in Lemma 1 we find



and thus

$$V_0 = \frac{z}{1-z}\ , \quad V_h = z + zV_h(1 + V_{h-1} + V_{h-1}V_{h-2} + \ldots + V_{h-1}\cdots V_0).$$

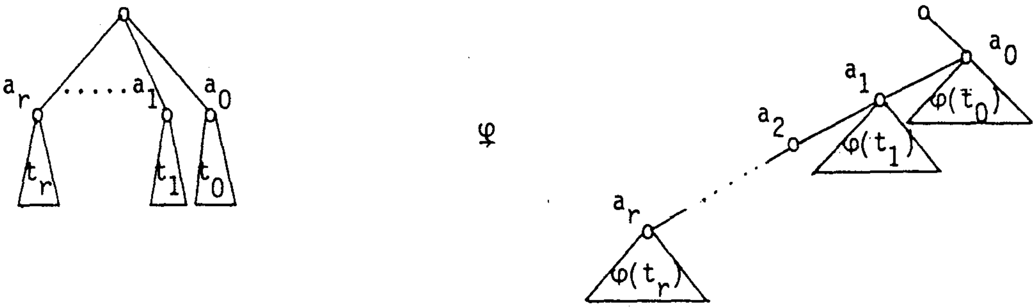
From this it is an easy induction to show that

$$V_0 = z/(1-z) \quad\text{and}\quad V_h = z/(1-V_{h-1}).$$

Since $V_0 = P_1$, a comparison with (2) finishes the proof.

We also present a proof by establishing a bijection $\varphi: V_h \to P_{h+1}$. The first step maps a tree with v-height $\leq h$ and n nodes onto a binary tree with h^*-height $\leq h$ and $n-1$ nodes. This is done recursively:

$$\circ \ \varphi\ \circ ,$$



Having performed this recursive operation, the root is to be deleted; this is the first step of our bijection. Regard that in fact φ is a version of the inverse of the "rotation correspondence" [7]. The second step is the classical version of this correspondence between binary trees with h^*-height $\leq h$ and $n-1$ nodes and planted plane trees with ordinary height $\leq h+1$ and n nodes. □

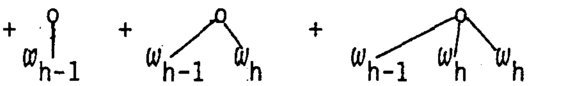
So the asymptotics of v_n are immediate from the asymptotics of h_n^* ([1]) and part (b) of the main theorem is proved.

We are now left with the proof of part (c) of the main theorem. While in the proofs of (a) and (b) our method was to establish an explicit connection with DE BRUIJN, KNUTH and RICE's result for the ordinary height of planted plane trees, another approach seems to be necessary to achieve (c). The more function theoretic approach was stimulated by the pioneering treatment of the problem of the average height of binary trees by FLAJOLET and ODLYZKO [3].

LEMMA 5. With $\varepsilon=\sqrt{1-4z}$ and $f_h(z) = y(z)-W_h(z)$,

$$f_h^2 + (\varepsilon+z)f_h - zf_{h-1} = 0.$$

Proof. We have

$\mathcal{W}_0$ = o and $\mathcal{W}_h$ = o +  + ... ,

whence

$$W_0 = z \text{ and } W_h = z + zW_{h-1}/(1-W_h)$$

from which the result follows by some easy manipulations. □

LEMMA 6. $\sum_{h\geq 0}(y-W_h) = -\frac{1}{8}\log\varepsilon + K + O(|1-4z|^{\nu})$ for $z \to \frac{1}{4}$ and for all $\nu < \frac{1}{4}$.

Proof. Because of the complexity of a complete treatment we omit the details and only stress the main steps:

Solving the quadratic equation of Lemma 5 and expanding the square root it follows that

$$f_h = \frac{z}{\varepsilon+z} f_{h-1} \left(1 - \frac{z}{(\varepsilon+z)^2} f_{h-1}\right) + \dots$$

With the substitution $g_h = \frac{z}{(\varepsilon+z)^2} f_h$,

$$g_h = \frac{z}{\varepsilon+z} g_{h-1} (1 - g_{h-1} + \dots) + \dots$$

Since $\frac{z}{\varepsilon+z} = 1 - 4\varepsilon + O(\varepsilon^2)$ it turns out that the behaviour of Σg_h is asymptotically equivalent to ΣG_h, with

$$G_h = (1-4\varepsilon)\, G_{h-1}\, (1 - G_{h-1}).$$

Adopting FLAJOLET and ODLYZKO's technique [3] it follows that

$$\sum_{h\geq 0} G_h = -\frac{1}{2}\log\varepsilon + K' + O(|\varepsilon|^{\nu}) \text{ for } z \to \frac{1}{4} \text{ and all } \nu < \frac{1}{4}$$

from which the lemma is obvious. □

Again making use of the "translation technique" cited above we finally arrive at part (c) of the main theorem.

We finish this section with some results related to the material from above.

Let $h_k(t)$ denote the maximal number of nodes of outdegree k in a chain connecting the root with a leaf. Furthermore let $H_{k,h}(z)$ be the generating function of the trees t with $h_k(t) \leq h$. Then we get

$$H_{k,h} = \frac{z}{1-H_{k,h}} - z\, H_{k,h}^{k} + z\, H_{k,h-1}^{k} \ .$$

With $e_{k,h}(z) = y(z) - H_{k,h}(z)$ we get in a similar way as above

$$e_{k,h} = e_{k,h-1}\left(1 - \frac{2^{k+2}}{k}\, e_{k,h-1}\right) + \ldots$$

and therefore

$$\sum_{h\geq 0} e_{k,h} = -\frac{k}{2^{k+3}}\log\varepsilon + K_k + O(|\varepsilon|^{\nu}),$$

so that the average value of the "height" $h_k(t)$ for trees t of size n is asymptotically equivalent to

$$\frac{k}{2^{k+1}}\sqrt{\pi n}\ . \tag{5}$$

A slightly different but related topic is now discussed: Following POLYA [8], resp. FORLINGER and HOFBAUER [5], we consider pairs of lattice paths in the plane, each path starting at the origin and consisting of unit horizontal and vertical steps in the positive direction.

Let $L_{n,j}$ be the set of such path-pairs (π,σ) with the following properties:

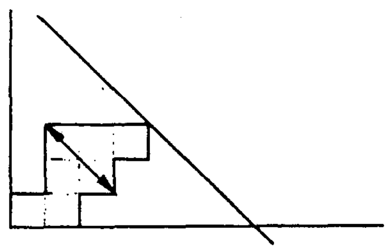
(i) both π and σ end at the point $(j,n-j)$

(ii) π begins with a unit vertical step and σ with a horizontal

(iii) π and σ do not meet between the origin and their common endpoint.

The elements of $L_n = \bigcup_{j=1}^{n} L_{n,j}$ are polygons with circumference 2n, and it is well known that $|L_n| = \frac{1}{n}\binom{2n-2}{n-1}$, $n \geq 2$; $|L_1| = 0$.

We define now the height $d(\pi,\sigma)$ of a path-pair (π,σ) to be the maximal length of a "diagonal" parallel to $y=-x$ between two lattice points on the path-pair, e.g.



has $d(\pi,\sigma) = 2$.

Let $D_h(z)$ denote the generating function of path-pairs (π,σ) with $d(\pi,\sigma) \le h$.

LEMMA 7. $D_h(z) = P_{2h}(z) - z$.

Proof. We use the bijection between L_n and "Catalan" words in $\{0,1\}^*$ described in [5]: Represent a path-pair $(\pi,\sigma) \in L_n$ as a sequence of pairs of steps: let v be a vertical step and h a horizontal step. The pair (π,σ) with $\pi = a_1 \dots a_n$, $\sigma = b_1 \dots b_n$ where each a_i and b_i is a v or h, is represented as the sequence of step-pairs $(a_1,b_1) \dots (a_n,b_n)$. To encode the sequence of step-pairs as a Catalan word the following translation is used:

$$(v,h) \to 00 \qquad (v,v) \to 10$$
$$(h,v) \to 11 \qquad (h,h) \to 01$$

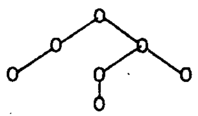
Omitting one "0" at the beginning and one "1" at the end a Catalan word is derived.

[For example: The path-pair (π,σ) from above is represented by the sequence

$$(v,h),(h,h),(v,v),(v,h),(h,v),(h,h),(h,v)$$

and encoded as the word 001100011011.]

The Catalan word is now represented in the well known way as a planted plane tree $t(\pi,\sigma)$ of size n.

[In the example  .]

We study now the influence of a step-pair (a_i,b_i) of the path-pair (π,σ) on the height of the corresponding nodes of the planted plane tree $t(\pi,\sigma)$:

If we had arrived at a node of height k before attaching the part of the tree corresponding to (a_i,b_i) the next two nodes will have heights

$$\begin{array}{rl} k-1,k & \text{if } (a_i,b_i) = (v,v) \leftrightarrow 10 \\ k+1,k+2 & \text{if } (a_i,b_i) = (v,h) \leftrightarrow 00 \\ k-1,k-2 & \text{if } (a_i,b_i) = (h,v) \leftrightarrow 11 \\ k+1,k & \text{if } (a_i,b_i) = (h,h) \leftrightarrow 01 \end{array} .$$

On the other hand the "local" diagonal distance l between the path-pairs develops as follows:

$$\begin{array}{rl} l & \text{if } (a_i,b_i) = (v,v) \leftrightarrow 10 \\ l+1 & \text{if } (a_i,b_i) = (v,h) \leftrightarrow 00 \\ l-1 & \text{if } (a_i,b_i) = (h,v) \leftrightarrow 11 \\ l & \text{if } (a_i,b_i) = (h,h) \leftrightarrow 01 \end{array} .$$

So it is an easy consequence that the set of all path-pairs (π,σ) with $d(\pi,\sigma) \le h$ corresponds to the set of trees t of size n with height of t equal to 2h-1 or 2h. Thus we have $D_h - D_{h-1} = P_{2h} - P_{2h-2}$, $h \ge 1$, with $D_0(z) = 0$. Summing up we get

$$D_h(z) = P_{2h}(z) - P_0(z) = P_{2h}(z) - z. \square$$

PROPOSITION. The average value of $d(\pi,\sigma)$ for path-pairs in L_n is

$$h_n - u_n = \frac{1}{2}\sqrt{\pi n} - \frac{1}{2} + O(n^{-1/2}).$$

Proof. Let $l(z) = \ddot{y}(z) - z$ denote the generating function of all path-pairs. Then, regarding Lemma 7 and Lemma 1,

$$\sum_{h\geq 0}(1-D_h) = \sum_{h\geq 0}(y-P_{2h}) = \sum_{h\geq 0}(y-P_h) - \sum_{h\geq 0}(y-U_h)$$

from which the result is immediate. □

In [5] there is another interesting bijection between path-pairs and planted plane trees. Let $(\pi,\sigma) \in L_{n,j}$ be a path-pair with steps $\pi = a_1\ldots a_n$, $\sigma = b_1\ldots b_n$ (a_i, b_i $\in\{v,h\}$). We decompose now π resp. σ in the following way:

For

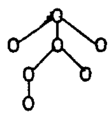
$$\pi = v^{s_1}\, h\, v^{s_2}\, h \ldots v^{s_j}\, h\,, \quad s_i \geq 0$$

$$\sigma = h\, v^{t_1}\, h\, v^{t_2} \ldots h\, v^{t_j}\,, \quad t_i \geq 0$$

we consider the "Catalan" word $\quad 0^{s_1}\, 1^{t_1+1}\, 0^{s_2+1}\, 1^{t_2+1} \ldots 1^{t_{j-1}+1}\, 0^{s_j+1}\, 1^{t_j}$

which again corresponds to a planted plane tree as usual.

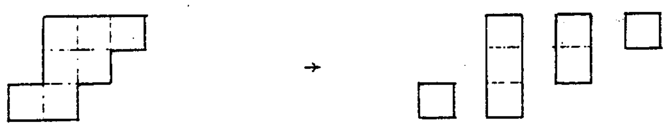
[In our example from above (π,σ) is encoded as 010001101101 and corresponds to

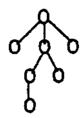


.]

It is easily seen that the height of the i-th leaf from the left of the tree constructed as indicated equals the area of the i-th vertical rectangle of width 1 from the left between π and σ.

[In our example the sequence of areas is 1,3,2,1, corresponding to



and 1,3,2,1 is also the sequence of heights of the leaves of the tree  .]

REFERENCES

[1] N.G. DE BRUIJN, D.E.KNUTH and S.O.RICE, The average height of planted plane trees in "Graph Theory and Computing" (R.C.Read ed.), 15-22, Academic Press, New York, 1972.

[2] P.FLAJOLET, Analyse d'algorithmes de manipulation d'arbres et de fichiers, Cahiers du BURO, 34-35 (1981), 1-209.

[3] P.FLAJOLET and A.ODLYZKO, The average height of binary trees and other simple trees, J.Comput.Syst.Sci. 25 (1982), 171-213.

[4] P.FLAJOLET, J.C.RAOULT and J.VUILLEMIN, The number of registers required to evaluate arithmetic expressions, Theoret.Comput.Sci. 9 (1979), 99-125.

[5] J.FÜRLINGER and J.HOFBAUER, q-Catalan numbers, preprint, Universität Wien, 1983.

[6] R.KEMP, The average number of registers needed to evaluate a binary tree optimally, Acta.Inf. 11 (1979), 363-372.

[7] D.E.KNUTH, "The Art of Computer Programming: Fundamental Algorithms", Addison-Wesley, Reading, Mass., 1968.

[8] G.POLYA, On the number of certain lattice polygons, J.Comb.Theory 6 (1969), 102-105.

[9] H.PRODINGER, The height of planted plane trees revisited, Ars Combinatoria, to appear, 1984.