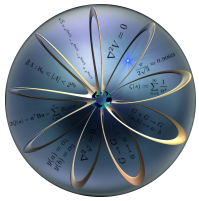

planetmath.org

Math for the people, by the people.

Dyck language

Canonical name	DyckLanguage
Date of creation	2013-03-22 18:55:25
Last modified on	2013-03-22 18:55:25
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	17
Author	CWoo (3771)
Entry type	Definition
Classification	msc 68Q45
Classification	msc 68Q42
Synonym	well-nested
Synonym	fully balanced
Synonym	parenthesis language
Related topic	DyckPaths
Related topic	ExampleOfCatalanNumbers
Defines	well-balanced

The importance of using parentheses can be illustrated by looking at the following expression:

$$((2-1)\cdot(-(1+2))\cdot 4)\div((1+2)\cdot 2)$$

There is no ambiguity in computing the result, which is -2. If we remove all the parentheses in the expression, we get

$$2-1\cdot -1+2\cdot 4\div 1+2\cdot 2$$

which does not make much sense, unless we know the order of arithmetic operations in advance. In addition, without using parentheses, the result will differ depending on how the order of operations is assigned.

Now, if we remove all the symbols in the first expression above except the parentheses, we get

$$(()(()))(())$$

an expression known as a word of "well-balanced" parentheses.

Formally, let $\Sigma_1 = \{(,)\}$ be an alphabet consisting of the left and right parentheses. Given word u over Σ_1, let $D_1(u)$ be the number of occurrences of the left parentheses in u minus the number of occurrences of the right parentheses in u.

Definition. A word u over Σ_1 is said to be a word of *well-balanced* parentheses, if

1. $D_1(u) = 0$, and
2. $D_1(v) \geq 0$ for any prefix v of u.

For simplicity, we also say that u is a well-balanced word over Σ_1.

Given this definition, the word above is well-balanced, but $()(()))$ and $)())(($ are not.

Definition. The set of well-balanced words over Σ_1 is called the *parenthesis language* or *Dyck language* over Σ_1, and is denoted by $\mathbf{Paren_1}$.

The 1 in Σ_1 denotes that only one type of parentheses is used in the language.

By induction, it is not hard to see that $\mathbf{Paren_1}$ can be generated by the following grammar:

1. terminal set is Σ_1,

2. non-terminal set is the singleton consisting of the start symbol σ,

3. productions are $\sigma \to \lambda$ (the empty word), $\sigma \to \sigma\sigma$, and $\sigma \to (\sigma)$.

As a result, $\mathbf{Paren_1}$ is context-free. Furthermore, $\mathbf{Paren_1}$ is a deterministic language, and hence unambiguous.

More generally, one can consider expressions involving more than one type of parentheses, such as [], {}, and $\langle\rangle$.

Definition. Let $\Sigma_n = \{(_1,)_1, \ldots, (_n,)_n\}$ be an alphabet consisting of n types of parentheses, a left and a right one for each type. The *Dyck language* over Σ_n, written $\mathbf{Paren_n}$, is the language generated by the following grammar:

1. terminal set is Σ_n,

2. non-terminal set is the singleton consisting of the start symbol σ,

3. productions are $\sigma \to \lambda$ (the empty word), $\sigma \to \sigma\sigma$, and $\sigma \to (_i\sigma)_i$ for each i in $\{1, \ldots, n\}$.

As before, $\mathbf{Paren_n}$ is context-free, and deterministic in particular, and hence unambiguous.

Words in $\mathbf{Paren_n}$ are also called *well-balanced*. However, it is a little more complicated to characterize what a well-balanced word is. The two criteria above for the case $n = 1$, while necessary, are not sufficient enough to describe the "well-nestedness" of parentheses when $n > 1$. For example, if $n = 2$, and the parentheses considered are {} and [], then the word [{]} satisfy both criteria above, but fail to be well-nested.

In order to fully characterize a well-balanced word over Σ_n, we first define, for each $i = 1, \ldots, n$, the function D_i much the same way as D_1: so that $D_i(u)$ is the number of left parentheses $(_i$, minus the number of right parentheses $)_i$. Call a word u partially balanced if, for every $i = 1, \ldots, n$:

1. $D_i(u) = 0$, and

2. $D_i(v) \geq 0$ for every prefix v of u.

Next, write $u = u_1 \cdots u_m$, where each u_k is a symbol in Σ_n. Let $u(j)$ be the prefix $u_1 \cdots u_j$. Given a position j in u, if u_j is a left parenthesis, say $(_i$, then there is a corresponding right parenthesis $)_i$ in u to the right of u_j, positioned at, say k, satisfying the equation $D_i(u(j)) = D_i(u(k)) + 1$. This

is a straightforward result of the two criteria above. Let j^+ be the least such position such that the equation holds. Now, if u_j is right parenthesis, then for some position $k < j$, we have $k^+ = j$. This means that, given any position j in u, there is a unique pair of positions (j_0, j_1), such that

- either $j = j_0$ or $j = j_1$, and
- $j_0^+ = j_1$.

Define, for each j, the word $u[j]$ to be the subword of u with starting position j_0 and ending position j_1. Now, we are ready to state the last criterion in order that u be well-balanced:

3. for each position j in u, the word $u[j]$ is partially balanced.

It can be shown, the set of words satisfying all three criteria above is $\mathbf{Paren_n}$. Furthermore, if $n = 1$, the third criterion can be derived from the first two criteria.

Other than being deterministic, some basic properties of $\mathbf{Paren_n}$:

- $\mathbf{Paren_m} \subseteq \mathbf{Paren_n}$, for any $m \leq n$.
- $\mathbf{Paren_n}$ is monoidal (it is a monoid): $\lambda \in \mathbf{Paren_n}$, and if $u, v \in \mathbf{Paren_n}$, then $uv \in \mathbf{Paren_n}$.
- More generally, $\mathbf{Paren_n}$ is insertion closed: if $u, vw \in \mathbf{Paren_n}$, then $vuw \in \mathbf{Paren_n}$.
- $\mathbf{Paren_n}$ is also deletion closed: if $u_1 v u_2, v \in \mathbf{Paren_n}$, then $u_1 u_2 \in \mathbf{Paren_n}$.
- $\mathbf{Paren_n}$ is not prefix-free.
- Suppose $f : \Sigma_n \to \Sigma_m^*$ is a function, such that for each $i = 1, \ldots, n$, f maps $(_i$ and $)_i$ to some u_i and v_i respectively such that $u_i v_i \in \mathbf{Paren_m}$. Then the extension $f^* : \Sigma_n^* \to \Sigma_m^*$, when restricted to $\mathbf{Paren_n}$, is a language homomorphism from $\mathbf{Paren_n}$ to $\mathbf{Paren_m}$.

Remark. It can be shown that the number of words of length $2n$ in $\mathbf{Paren_1}$ is the n-th Catalan number. For a proof of this, see `http://planetmath.org/ExampleOfCa` entry. The idea is to visualize a word in $\mathbf{Paren_1}$ as a path in a two-dimensional lattice, which can be generated as follows: given a word u of

length $2n$, the path $p(u)$ starts from $(0,0)$ (which corresponds to the first symbol of u). If point (i,j) is on the path, then the next point on the path is $(i+1,k)$, where $k=j+1$ if the i-th symbol of u is (, otherwise $k=j-1$. So the increase from one point to the next is either $(1,1)$, or $(1,-1)$. As a result, the path $P(u)$ has the property that it never dips below the x-axis, and it ends at $(2n,0)$. Paths defined this way are also known as *Dyck paths*.

References

[1] D. C. Kozen, *Automata and Computability*, Springer, New York (1997).

[2] J.E. Hopcroft, J.D. Ullman, *Formal Languages and Their Relation to Automata*, Addison-Wesley, (1969).