Tail Symbols

The (current ?) definition Quoting the paper : 'we call tail symbol the symbol that follows the last symbol inserted in the DST'. What I'm wondering is if the tail symbol is picked from

- the same sequence that generated the phrase that was just inserted into the DST
- or the beginning of the sequence that follows this phrase ?

In the latter case, I'm considering the same example as in the paper : the parsing of the eight sequences (these are only their inserted prefixes) :

$$1 \quad 10 \quad 0 \quad 101 \quad 00 \quad 01 \quad 000 \quad 100$$

Is this figure then accurate about the tail symbols ? (I hope not because this definition seems to contradict other things)

$$\begin{tikzcd}
(1) \arrow[r, bend right, "\text{tail} = 1"'] & (10) \arrow[r, bend right, "\text{tail} = 0"'] & (0) \arrow[r, bend right, "\text{tail} = 1"'] & (101) \arrow[r, bend right, "\text{tail} = 0"'] & (00) \arrow[r, bend right, "\text{tail} = 0"'] & (01) \arrow[r, bend right, "\text{tail} = 0"'] & (000) \arrow[r, bend right, "\text{tail} = 1"'] & (100)
\end{tikzcd}$$

Using this definition

- 0 occurs four times as the tail symbol
- 1 occurs three times

Therefore, in the last paragraph before equation (1), $\mathrm{T}_8{}^{(0)}$ should be 4 rather than 5 ? And we could now define the tail symbols as *the first symbol of each sequence except the first one* ?

But then it seems to clash with the definition of $\mathrm{T}_n = (\mathrm{T}_n^a, \mathrm{T}_n^b)$, having T_n^a as the number of times 'a' appears as a tail symbol assuming that *all* sequences start with symbol 'a'. If *all* means 'all the n sequences' then the tail symbols definition is absurd because all the tail symbols are just 'a' ?

As for the recursion

$$\mathrm{T}_{n+1}^a = \delta_a + \mathrm{T}_{n_a}^a + \mathrm{T}_{n_b}^b$$

δ_a is "equal to 1 when the second symbol of the first sequence is 'a'" suggests that we should use all the first sequence and not only its prefix that is being inserted in the DST ?

Another final problem with this definition is that I don't see the probabilistic value of the first symbol of each independent sequence : I could actually initialize them with any value, although I'd probably use the stationary distribution to pick that first symbol.

The (right ?) definition Therefore I think right now that the definition of the tail symbols would rather be, since we are in the Markov Independent model : if a phrase

$$p = w_1 \ldots w_n$$

is inserted from a sequence

$$\mathrm{X} = w_1 \ldots w_n \, w_{n+1} \ldots$$

then the tail symbol of that phrase is w_{n+1}. Therefore, to define the tail symbols, for example for $n = 4$, with the sequences $\mathrm{X}(1) = 00000\ldots$, $\mathrm{X}(2) = 1010101\ldots$, $\mathrm{X}(3) = 1001101\ldots$ and $\mathrm{X}(4) = 001100111\ldots$, which give the parsing

$$()(0)(1)(10)(00)$$

we would do, with the prefix phrases in red and the tail symbol in green :

$$\begin{aligned} \mathrm{X}(1) &= 0000000\ldots \\ \mathrm{X}(2) &= 1010101\ldots \\ \mathrm{X}(3) &= 1001101\ldots \\ \mathrm{X}(4) &= 001100111\ldots \end{aligned}$$

This means that the tail symbols are not obvious from the DST, and this reconciles with the definition of δ_a. If this is the correct definition, then the tree of figure 1 is misleading because we need the sequences to define the tail symbols. (as well as the phrase 'in Figure 1 the tail symbol after phrase (10) is "0")