Parkh theorem for data grammars

Let Σ be a finite alphabet. Parikh image is a homomorphism that goes from words over Σ with the concatenation operation i.e. Σ^* into functions $\Sigma \to \mathbb{N}$ with pointwise addition. The Parikh image of a word is simply a function that for every later $a \in \Sigma$ returns the number of occurrences of that later in the word. The well-known Parikh's theorem says that for every context-free grammar there exists a finite automaton such that Parikh images of their languages are the same.

The question that we ask is about a generalization of the Parikh theorem. Let $\mathbb{D}$ be an infinite unordered data set. You should think of it as about the set of strings or barcodes. For the remaining part of the problem statement we identify $\mathbb{D}$ with the set of natural numbers $\mathbb{N}$.

A data grammar with k registers is a quadruple $(N, \Sigma, S(1, \dots k), R)$ where:

- N is a finite set of nonterminals of arity k,
- Σ is a set of terminal symbols of arity 1,
- $S(1, 2, 3 \dots k)$ is an initial nonterminal symbol with the initial valuation,
- R is a set of rules of the following form:

$$X(\alpha_1, \alpha_2 \dots \alpha_k) \to Y(\beta_1, \beta_2 \dots \beta_k)Z(\gamma_1, \gamma_2 \dots \gamma_k)$$
$$X(\alpha_1, \alpha_2 \dots \alpha_k) \to a(\beta_1)$$
$$X(\alpha_1, \alpha_2 \dots \alpha_k) \to \epsilon$$

where ϵ is the empty word, $a(\beta_1)$ is an arity 1 terminal symbol, and X, Y, Z are k-arity nonterminals. $\alpha_i, \beta_i, \gamma_i$ are variables not necessary pairwise different.

To understand how we apply rules let us consider example: Suppose we have a configuration $X(1, 3, 4)b(4)X(2, 2, 3)Y(2, 2, 2)$ and we apply a rule $X(\alpha_1, \alpha_2, \alpha_3) \to X(\alpha_1, \alpha_1, \alpha_1)Y(\alpha_2, \alpha_3, \alpha_4)$. To apply this rule α_1 has to be evaluated to 1, α_2 to 3, and α_3 to 4. α_4 can be evaluated to anything different than $1, 3, 4$, say to 6. Then we get simply $X(1, 1, 1)Y(3, 4, 6)b(4)X(2, 2, 3)Y(2, 2, 2)$.

The analogues of finite automata(called register automata) are data grammars such that for every rule there can be at most one nonterminal on the right hand side and it has to be the last symbol in this rule, like below.

$$X(\alpha_1, \alpha_2 \dots \alpha_k) \to a(\beta_1)Z(\gamma_1, \gamma_2 \dots \gamma_k)$$

Data grammars are recognizing words in $\Sigma(\mathbb{D})^*$.

Parikh image for data words is a function from $\Sigma(\mathbb{D})^* \to (\Sigma(\mathbb{D}) \to \mathbb{N})$ defined similarly to normal Parikh image.

Now the natural question is if the sets of Parikh images of languages generated by data grammars with k registers and Parikh images of languages recognized by register automata with k registers are the same.

The hypothesis holds for $k = 1$, what about $k = 2, 3 \dots$?