A fibrational characterization for unicity of solutions to generalized context-free systems

Farzad Jafarrahmani
LIP6, Sorbonne Université

Noam Zeilberger
LIX, Ecole Polytechnique

In this work we develop a categorical perspective on the question of when a context-free grammar, considered as a system of recursive equations on the languages generated by its non-terminals, has a unique solution. We arrived at this question in the context of a more open-ended project to develop a fibrational semantics of cyclic proofs, building on an analogy between cyclic proofs and finite-state automata under which the subderivations of a proof correspond to the states of a machine, and inference rules to transitions. Both context-free grammars and finite-state automata are typically "circular", in the sense that the production rules of the grammar and the transitions of the automaton most often induce cycles.

Recently, Melliès and Zeilberger [3, 4] proposed categorical abstractions of the notions of context-free grammar and of non-deterministic finite state automaton as certain kinds of functors between categories and operads, under which a **generalized context-free grammar** (gCFG) is defined as

a functor from a free operad generated by a pointed finite species into an arbitrary operad

while a **generalized non-deterministic finite-state automaton** (gNDFA) is defined as

a functor satisfying the unique lifting of factorizations and finite fiber properties.

These definitions permit for example to give a clean categorical proof that context-free languages are closed under intersection with regular languages. Here we focus on generalized context-free grammars, and consider a related view under which a gCFG is seen as inducing a recursive system of polynomial equations, which may be interpreted fibrationally.

It is best to illustrate with an example. Consider the following CFGs with labelled production rules:

$$G_1 \;=\; \begin{array}{lll} S & \to^a & \varepsilon \\ S & \to^b & [S] \\ S & \to^c & SS \end{array} \qquad G_2 \;=\; \begin{array}{lll} S & \to^d & \varepsilon \\ S & \to^e & [S]S \end{array} \tag{1}$$

Both grammars generate the same language $L = \mathcal{L}_{G_1} = \mathcal{L}_{G_2}$, namely the Dyck language of balanced brackets. However, they may be seen as implicitly stating two different equations *satisfied* by this language,

$$L = \varepsilon + [L] + LL \tag{2}$$

$$L = \varepsilon + [L]L \tag{3}$$

where we have written $+$ for union of languages. As we will explain, such equations can be interpreted as properties of the initial models of the grammars G_1 and G_2 in the "refinement system" Subset $\to$ Set of subsets over sets. At a more elementary level, what we want to emphasize here is that although it is easy to show that the Dyck language is the *minimal* solution to both equations, equation (2) has many solutions (e.g., $L = \Sigma^*$), while equation (3) has a *unique* solution.

Submitted to:

Indeed, there are at least two different ways of interpreting a context-free language as the solution to a recursive system of constraints. Under the traditional view, the language generated by a grammar is defined *inductively* as the smallest language closed under the production rules. This corresponds to interpreting the languages L_1 and L_2 generated by G_1 and G_2 as the smallest languages $L_1 = \mu(F_1), L_2 = \mu(F_2)$ satisfying the respective inclusions $F_1(L_1) \subseteq L_1$ and $F_2(L_2) \subseteq L_2$, where the operators $F_1, F_2 : P(\Sigma^*) \to P(\Sigma^*)$ are defined by

$$F_1(X) = \varepsilon + [X] + XX \qquad F_2(X) = \varepsilon + [X]X$$

On the other hand, one can take the recursive equations (2) and (3) literally and consider their solutions. In this case, (2) is inadequate for determining the language, but (3) is a perfectly valid (even if circular) definition since it has a unique solution.

Intuitively, the reason why (3) has a unique solution is because the rule $S \to^e [S]S$ "consumes" a pair of letters, so it is not possible to build an infinite derivation of a finite word in G_2. On the other hand, it is possible to iterate rules a and c of G_1 to build a "vicious cycle"

$$S \to^c SS \to^a S \to^c SS \to^a S \ldots$$

which may be seen as an ill-founded derivation of an arbitrary word, hence explaining why $L = \Sigma^*$ is another solution to (2).

It is possible to formulate the question of unicity of solutions to equations arising from gCFGs in a very general fibrational framework (adapted from [4, Addendum A]), as a question about interpretations

$$\begin{array}{ccc} \mathsf{Free}(\mathcal{S}) & \xrightarrow{\tilde{M}} & \mathcal{E} \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle q} \\ \mathcal{O} & \xrightarrow{M} & \mathcal{B} \end{array}$$

of a gCFG p in an arbitrary functor q that is "polynomially closed", in the sense that it admits push-forwards and fiberwise coproducts (needed for interpreting equations such as (2) and (3)). One benefit of doing so is that we can consider "proof-relevant" models of gCFGs in $\mathrm{Set}^{\to} \to \mathrm{Set}$, under which a language is interpreted not merely as a subset of words but as an assignment of a family of derivations to every word. We focus on the question of unicity of solutions in the proof-relevant model $\mathrm{Set}^{\to} \to \mathrm{Set}$, since it implies unicity of solutions in the proof-irrelevant model $\mathrm{Subset} \to \mathrm{Set}$.

After several attempts at devising conditions for unicity which were sufficient but far from necessary, we finally arrived at a condition that we call *relative nilpotency*. The starting point is to consider the base operad $\mathcal{O}$ as being equipped with a *non-unital suboperad* $\mathcal{O}^+ \subset \mathcal{O}$, whose operations induce a well-founded ordering on the constants of $\mathcal{O}$. In the case of classical CFGs, $\mathcal{O}$ is the *operad of spliced words* $\mathcal{W}[\Sigma]$ whose n-ary operations are sequences of $n+1$ words $w_0 - \ldots - w_n$, while $\mathcal{O}^+$ is its non-unital suboperad $\mathcal{W}[\Sigma]^+$ whose n-ary operations are sequences of $n+1$ words containing at least one non-empty word. This non-unital suboperad induces a well-founded ordering on the constants of $\mathcal{W}[\Sigma]$, namely on words u, where we declare that $u \succ u_1, \ldots, u \succ u_n$ if there exists an operation $f = w_0 - \ldots - w_n$ of $\mathcal{W}[\Sigma]^+$ such that $u = f(u_1, \ldots, u_n)$.

To formulate the relative nilpotency condition, we make use of the *composition product* from the theory of species [2], suitably adapted to colored, non-symmetric species (cf. [4, §1.6]). In this setting, the composition $\mathcal{S} \circ \mathcal{R}$ of two species with the same underlying set of colors (= non-terminals) is the species whose n-ary nodes $R_1, \ldots, R_n \to S$ consist of formal compositions $f \bullet (g_1, \ldots, g_k)$ of a k-ary node

$g : S_1, \ldots, S_k \to S$ of $\mathcal{S}$ with a k-tuple of nodes $f_1 : \Gamma_1 \to S_1, \ldots, f_k : \Gamma_k \to S_k$ of $\mathcal{R}$, such that $\Gamma_1, \ldots, \Gamma_k = R_1, \ldots, R_n$. Note that the composition product has a *unit* given by the species $\mathbb{I}$ with a single unary node $*_R : R \to R$ for every color R. We also write $\mathcal{R}^-$ for the species $\mathcal{R}^- := \mathcal{R} - \mathcal{R}(0)$ obtained by removing all nullary nodes from any species $\mathcal{R}$. Finally, we write $\Delta_{\mathcal{S}}$ for the endofunctor $\Delta_{\mathcal{S}} : \mathrm{Spec}_X \to \mathrm{Spec}_X$ on the category of species (with same underlying set of colors X as $\mathcal{S}$) defined by $\Delta_{\mathcal{S}} := \mathcal{R} \mapsto (\mathcal{R} \circ \mathcal{S})^-$. Then the **relative nilpotency** condition on a gCFG $p : \mathsf{Free}(\mathcal{S}) \to \mathcal{O}$ states that

$$\textit{there exists a } k \textit{ such that } p(\Delta_{\mathcal{S}}^k \mathbb{I}) \subset \mathcal{O}^+.$$

Observe that grammar G_1 above does not satisfy the relative nilpotency condition, while G_2 satisfies it with $k = 1$. We call this "relative" nilpotency because in the special case where $\mathcal{O} = 1$ is the terminal operad (which has a single color and a single n-ary operation of every arity n), the non-unital suboperad is trivial $\mathcal{O}^+ = 0$, and the condition reduces to the requirement that there exists a k such that $\Delta_{\mathcal{S}}^k \mathbb{I} = 0$, which is equivalent to asking that the grammar has no transitive cycles and therefore generates a finite language.

We note that the problem of characterizing unicity of solution to systems of polynomial equations induced by context-free grammars was considered in early work of Courcelle,[1] and that our relative nilpotency condition is very similar to one of the necessary and sufficient conditions he states as Proposition 15.10 of [1]. Indeed, our result may be seen as a modern categorical formulation and generalization of Courcelle's. It is worth mentioning that Courcelle [1] developed a broader "unified theory" about recursive definitions, and likewise, we eventually want to deal with other examples including cyclic proofs as well as recursive definitions in type theory and functional programming. That is one of our main motivations for using the unifying language of category theory. We were also inspired by Joyal's Implicit Species Theorem [2, Thm. 6], which answers a similar question of when a species is uniquely determined by a system of recursive equations.

References

[1] Bruno Courcelle (1986): *Equivalences and transformations of regular systems–applications to recursive program schemes and grammars*. Theoretical Computer Science 42, pp. 1–122.

[2] André Joyal (1981): *Une théorie combinatoire des séries formelles*. Advances in Mathematics 42(1), pp. 1–82.

[3] Paul-André Melliès & Noam Zeilberger (2022): *Parsing as a lifting problem and the Chomsky-Schützenberger representation theorem*. In: MFPS 2022 - 38th conference on Mathematical Foundations for Programming Semantics, doi:10.46298/entics.10508.

[4] Paul-André Melliès & Noam Zeilberger (2023): *The categorical contours of the Chomsky-Schützenberger representation theorem*. Available at `https://hal.science/hal-04399404`.

[1] We gratefully thank Sylvain Salvati for pointing us to this work.