Generalizing and abstracting the notion of context-free language

Noam Zeilberger[1]

Ecole Polytechnique

Lambda Pros day
Paris, 28 June 2023

[1] Joint work with Paul-André Melliès

Introduction: context-free languages of arrows

CFG over a category

In "Parsing as a lifting problem and the Chomsky-Schützenberger representation theorem" (MFPS 2022), we proposed a definition of *context-free grammar over a category*.

- A category $\mathcal{C}$
- A finite species $\mathcal{S}$
- A functor $p : \mathsf{F}\,\mathcal{S} \to \mathsf{W}\,\mathcal{C}$
- A distinguished color $S \in \mathcal{S}$

where $\mathsf{F}\,\mathcal{S}$ is the free operad generated by $\mathcal{S}$, and where $\mathsf{W}\,\mathcal{C}$ is the operad of *spliced arrows in* $\mathcal{C}$.

The spliced arrow operad $\mathsf{W}\mathcal{C}$

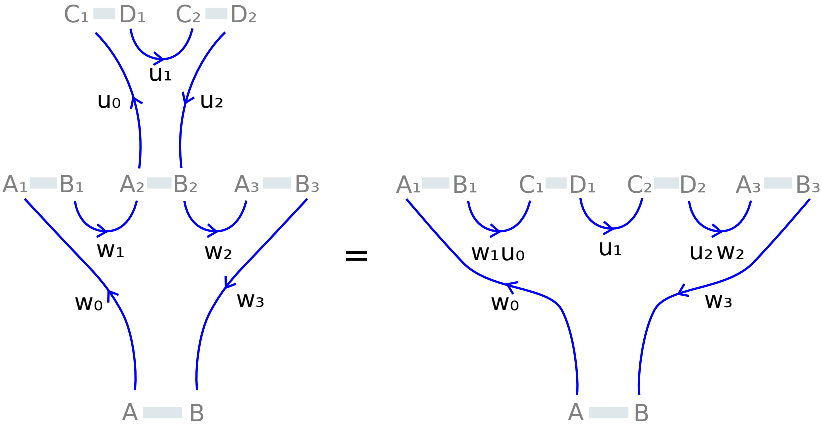
Colors are pairs (A, B) of objects of $\mathcal{C}$.

Operations $w_0 - w_1 - \ldots - w_n : (A_1, B_1), \ldots, (A_n, B_n) \to (A, B)$ consist of sequences of $n+1$ arrows in $\mathcal{C}$, where $w_i : B_i \to A_{i+1}$ for $0 \leq i \leq n$ under the convention that $B_0 = A$ and $A_{n+1} = B$.

The identity operation on (A, B) is given by $id_A - id_B$.

Composition performed by "splicing into the gaps" (see next slide).

The spliced arrow operad $W\mathcal{C}$



The spliced arrow operad $\mathsf{W}\mathcal{C}$

The spliced arrow operad construction has a left adjoint, which we called the "contour category" of an operad.

$$\mathrm{Cat} \underset{\mathsf{W}}{\overset{\mathsf{C}}{\underset{\longrightarrow}{\overset{\longleftarrow}{\bot}}}} \mathrm{Operad}$$

This adjunction is fundamental to our analysis of the C-S theorem, but I won't use it in the talk. (See the MFPS paper for details.)

The language of arrows generated by a grammar

Let $G = (\mathcal{C}, \mathcal{S}, p, S)$. The **language of arrows** of G is the subset

$$L_G = \{ p(\alpha) \mid \alpha : S \} \subseteq \mathcal{C}(A, B)$$

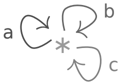
where $p(S) = (A, B)$.[2]

For example, any CFL in the classical sense is the language of arrows of a CFG over a one-object category B_Σ freely generated by an arrow $a : * \to *$ for every letter $a \in \Sigma$ of the alphabet.

[2] Which we often write as $S \sqsubset (A, B)$, saying that S *refines* the type (A, B).

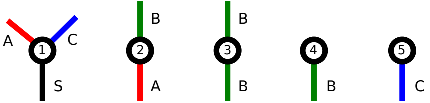
Example

$\mathcal{C}$

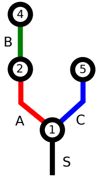



```
1 : S → AaCc
2 : A → aB
3 : B → aBb
4 : B → b
5 : C → c
```

$\mathcal{S}$



F$\mathcal{S}$



↧ ↧ ↧ ↧ ↧ ↧

W$\mathcal{C}$

ε-a-c	a-ε	a-b	b	c	ε-a-c ∘ (a-ε∘b,c) = abacc

Motivations

Some motivations for modelling CFGs as functors $p : \mathsf{F}\,\mathcal{S} \to \mathsf{W}\,\mathcal{C}$

- Builds on our work modelling type systems as functors
- Can reformulate many standard properties more simply
- Parsing becomes a *lifting problem* along the functor p

Some motivations for CFGs over proper categories (> 1 object)

- Typed words $w : A \to B$ yield a more elegant implementation of common parsing hacks, such as an end-of-input symbol \$.
- Can take the *pullback* of a CFG along an NDFA over the same category, to define a CFG over the automaton! The usual intersection construction is thereby decomposed in two steps.

This talk[3]

Further generalize and abstract the notions of CFG and CFL:

1. Define *generalized CFGs* replacing $\mathsf{W}\,\mathcal{C}$ by arbitrary operad $\mathcal{O}$.
2. Redefine CFLs as *initial models* of CFGs, for an appropriate notion of model.

Why (1)? It's mathematically natural, and allows us to cover interesting examples from the literature.

Why (2)? It formalizes the old idea that CFLs may be viewed as minimal solutions to systems of polynomial equations, while also allowing us to incorporate "proof-relevant" languages.

[3]Based on work-in-progress, not in the MFPS version of the paper.

Generalized context-free grammars

CFG over an operad

A **generalized CFG** $G = (\mathcal{O}, \mathcal{S}, p, S)$ is given by

- An operad $\mathcal{O}$
- A finite species $\mathcal{S}$
- A functor $p : \mathsf{F}\,\mathcal{S} \to \mathcal{O}$
- A distinguished color $S \in \mathcal{S}$

The *language of constants* generated by G is the subset

$$L_G = \{\, p(\alpha) \mid \alpha : S \,\} \subseteq \mathcal{O}(A)$$

where $S \sqsubset A$.

Example: multiple & parallel CFGs (Seki et al., 1991)

For any operad $\mathcal{P}$, one can build operads $L_{\mathrm{sym}}\,\mathcal{P}$ / $L_{\mathrm{aff}}\,\mathcal{P}$ / $L_{\mathrm{cart}}\,\mathcal{P}$:

- colors are lists $[A_1, \ldots, A_k]$ of colors of $\mathcal{P}$
- operations

  $$([f_1, \ldots, f_k], \sigma) : [\Gamma_1], \ldots, [\Gamma_n] \to [A_1, \ldots, A_k]$$

  are given by a pair of a list of operations

  $$f_1 : \Omega_1 \to A_1, \ldots, f_k : \Omega_k \to A_k$$

  of $\mathcal{P}$ together with a bijection / injection / function $\sigma : \Omega_1, \ldots, \Omega_k \to \Gamma_1, \ldots, \Gamma_n$.

These are, respectively, the free symmetric / semi-cartesian (or "affine") / cartesian monoidal operads over $\mathcal{P}$.

Example: multiple & parallel CFGs (Seki et al., 1991)

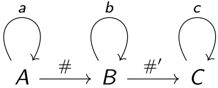
Observe that if $\mathcal{P}$ is an un(i)colored operad, then the colors of $L_{\mathsf{sym}}\,\mathcal{P}/L_{\mathsf{aff}}\,\mathcal{P}/L_{\mathsf{cart}}\,\mathcal{P}$ are simply (isomorphic to) natural numbers.

A gCFG over $L_{\mathsf{aff}}\,\mathsf{W}\,B_\Sigma$ with start symbol $S \sqsubset 1$ is precisely a **multiple context-free grammar** à la Seki et al. More generally, a gCFG over $L_{\mathsf{aff}}\,\mathsf{W}\,\mathcal{C}$ could be called a "multiple CFG of arrows".

Such a grammar is a *k-multiple CFG* just in case every non-terminal refines a list of length $\leq k$.

For **parallel multiple** CFGs, just replace $L_{\mathsf{aff}}\,\mathcal{P}$ by $L_{\mathsf{cart}}\,\mathcal{P}$.

Example: multiple & parallel CFGs (Seki et al., 1991)

We can define a 3-mCFG over the category 

generating the language $a^n \# b^n \#' c^n$, with two colors

$$S \sqsubset [(A, C)] \qquad R \sqsubset [(A, A), (B, B), (C, C)]$$

and a triple of operations in $\mathcal{S}$

$$x_1 : R \qquad x_2 : R \to R \qquad x_3 : R \to S$$

mapped respectively to the following operations in $L_{\text{aff}}\,\mathsf{W}\,\mathcal{C}$

$$([id_A, id_B, id_C], id) \quad ([a{-}id_A, b{-}id_B, c{-}id_C], id) \quad ([-\#-\#'-], id)$$

Example: series-parallel graphs (Courcelle & Engelfriet, 2012)

We can define a gCFG over the (large) operad Set, generating the set of series-parallel graphs:

- $\mathcal{S}$ has one color S which is mapped to the set $\mathrm{DiGr}_{\bullet,\bullet}$ of finite directed graphs with two distinct marked vertices.
- $\mathcal{S}$ has a pair of binary operations

  $$par, seq : S, S \to S$$

  mapped respectively to the operations

  $$(\|), (;) : \mathrm{DiGr}_{\bullet} \times \mathrm{DiGr}_{\bullet} \to \mathrm{DiGr}_{\bullet}$$

  of *parallel composition* and *series composition* of marked digraphs, as well as a constant $e : S$ mapped to the digraph $\bullet \to \bullet$ with one edge and two vertices.

Closure properties of classical CFLs

Union: if $L_1, L_2 \subseteq \Sigma^*$ are CF, then so is $L_1 \cup L_2 \subseteq \Sigma^*$

Concatenation: if $L_1, \ldots, L_n \subseteq \Sigma^*$ are CF, so is $L_1 \cdots L_n \subseteq \Sigma^*$

Homomorphic image: if $L \subseteq \Sigma^*$ is CF and $\phi : \Sigma^* \to \Pi^*$ is a monoid homomorphism, then $\phi(L) \subseteq \Pi^*$ is CF

Intersection with regular languages: if $L \subseteq \Sigma^*$ is CF and $R \subseteq \Sigma^*$ is regular, then $L \cap R \subseteq \Sigma^*$ is CF

Closure properties of generalized CFLs

Union: if $L_1, L_2 \subseteq \mathcal{O}(A)$ are CF, then so is $L_1 \cup L_2 \subseteq \mathcal{O}(A)$

Combination: if $L_1 \subseteq \mathcal{O}(A_1), \dots, L_n \subseteq \mathcal{O}(A_n)$ are CF, and $f : A_1, \dots, A_n \to B$ an op of $\mathcal{O}$, then $f(L_1, \dots, L_n) \subseteq \mathcal{O}(B)$ is CF

Functorial image: if $L \subseteq \mathcal{O}(A)$ is CF and $F : \mathcal{O} \to \mathcal{P}$ is a functor of operads, then $F(L) \subseteq \mathcal{P}(F\,A)$ is CF

Intersection with regular languages: if $L \subseteq \mathcal{O}(A)$ is CF and $R \subseteq \mathcal{O}(A)$ is regular[4], then $L \cap R \subseteq \mathcal{O}(A)$ is regular.

[4]We say that a language of constants is regular if it is recognized by an operadic NDFA = it is the image of some color $q \in \mathcal{Q}$ along a *finitary ULF* functor of operads $\mathcal{Q} \to \mathcal{O}$. Regular word languages and regular tree languages are recovered as special cases. As previously alluded to, intersection closure reduces to a more fundamental closure of gCFLs under pullback along NDFAs, combined with functorial image.

gCFLs as initial models of gCFGs

CFLs as minimal solutions to polynomial equations

Consider two different grammars for well-bracketed words:

$$G_1 = \begin{array}{lcl} S & \to & \epsilon \\ S & \to & [S] \\ S & \to & SS \end{array} \qquad G_2 = \begin{array}{lcl} S & \to & \epsilon \\ S & \to & [S]S \end{array}$$

Although the language $WB = L_{G_1} = L_{G_2}$ generated by both grammars is the same, G_1 and G_2 may be seen as implicitly stating two different equations *satisfied* by this language:

$$L = \epsilon + [L] + LL \quad (1)$$

$$L = \epsilon + [L]L \quad (2)$$

WB is the *minimal* solution to (1) in the sense it is contained in any language L such that $L = \epsilon + [L] + LL$. It is also the minimal solution[5] to (2).

[5] In fact $L = \epsilon + [L]L$ has a *unique* solution, for somewhat special reasons...

CFLs as minimal solutions to polynomial equations

An idea first formalized by Ginsburg & Rice (1962), further developed by Mezei & Wright (1967).

Also advocated by John Conway in his textbook (1971):

> In the standard treatment [of context-free languages] the transient letters are construction letters used as scaffolding in forming the language, but then discarded. We propose to develop the theory in a less orthodox way, in which this scaffolding never appears. We directly characterize the terminal images of the transient letters in terms of certain equations they satisfy.
>
> *Regular Algebra and Finite Machines,* Ch. 10, p. 80

gCFGs as sketches, gCFLs as initial models

Rather than do away with the scaffolding (as per Conway), we will treat a gCFG as defining a certain kind of "sketch"[6], which induces a category of models in some target space. gCFLs are then defined as *initial* models of gCFGs.

To make this precise, we first need to introduce some fibrational concepts for functors of operads, which will categorify systems of polynomial equations.

[6] In the spirit of Ehresmann, and formally very similar to the sketches used by Shulman in "LNL polycategories and doctrines of linear logic" (LMCS 19:2).

Notation

Given a functor of operads $q : \mathcal{E} \to \mathcal{B}$, we write

$$\Omega \overset{q}{\sqsubset} \Delta$$

to indicate Ω is a list of colors in $\mathcal{E}$ with image Δ in $\mathcal{B}$, and

$$\alpha : R_1, \ldots, R_n \underset{f}{\overset{q}{\Longrightarrow}} R$$

to indicate that $\alpha : R_1, \ldots, R_n \to R$ is an operation in $\mathcal{E}$ with image f in $\mathcal{B}$. We sometimes omit q when clear from context.

We also write $\mathcal{E}_f(R_1, \ldots, R_n; R)$ for the set of operations

$$\mathcal{E}_f(R_1, \ldots, R_n; R) = \{\, \alpha \mid \alpha : R_1, \ldots, R_n \underset{f}{\overset{q}{\Longrightarrow}} R \,\}.$$

Minimal cones

A **cone** in an operad $\mathcal{O}$ is a family of operations $(g_i : \Delta_i \to A)_{i \in I}$ in $\mathcal{O}$ with the same output color A.

Let $q : \mathcal{E} \to \mathcal{B}$ be a functor of operads. A cone $(\alpha_i : \Omega_i \Rightarrow^q_{g_i} R)_{i \in I}$ in $\mathcal{E}$ is said to be **minimal** over a cone $(g_i : \Delta_i \to A)_{i \in I}$ in $\mathcal{B}$ (relative to q) if for every operation $f : \Gamma, A, \Gamma' \to B$ in $\mathcal{B}$ with $|\Gamma| = k$, the function

$$(- \circ_k \alpha_i)_{i \in I} \quad : \quad \mathcal{E}_f(\Theta, R, \Theta'; S) \longrightarrow \prod_{i \in I} \mathcal{E}_{f \circ_k g_i}(\Theta, \Omega_i, \Theta'; S)$$

induced by precomposition with the α_i is invertible.

Given $(g_i : \Gamma_i \to A)_{i \in I}$ and $(\Omega_i \sqsubset^q \Gamma_i)_{i \in I}$, there exists at most one *q-minimal lift* of $(g_i)_i$ to $(\Omega_i)_i$, up to canonical isomorphism.

Special case: pushforward

A single operation $\alpha : \Omega \Rightarrow_g^q R$ of $\mathcal{E}$ is a minimal cone just in case it is (strongly) opcartesian relative to the functor of operads q.[7] In this case, we say R is the **pushforward** of Ω along g, generalizing the act of taking the image of a subset along a function.

[7] See Hermida (2000, 2004) for this notion, which extends the classical notion of opcartesian arrow relative to a functor of categories.

Special case: fiberwise coproduct

A cone $(\alpha_i : R_i \Rightarrow_{id_B} R)_{i \in I}$ of operations in $\mathcal{E}$ all lying over the same identity operation in $\mathcal{B}$ is minimal just in case R is the **fiberwise coproduct** of the R_i, generalizing the act of taking the union of subsets of a set. This means in particular that we have

$$\mathcal{E}_f(\Theta, R, \Theta'; S) \cong \prod_{i \in I} \mathcal{E}_f(\Theta, R_i, \Theta'; S)$$

for every compatible operation f.

General case

We write $\sum_{i\in I} g_i\,\Omega_i$ for some choice of object R coming together with a minimal cone $\left(in_j : \Omega_j \Rightarrow_{g_j} \sum_{i\in I} g_i\,\Omega_i\right)_{j\in I}$.

Proposition

Let $q : \mathcal{E} \to \mathcal{B}$ be a functor of operads. TFAE:[8]

1. *There is a minimal lift $\sum_{i\in I} g_i\,\Omega_i \sqsubset A$ of every cone $(g_i : \Gamma_i \to A)_{i\in I}$ in $\mathcal{B}$ to any family $\Omega_i \sqsubset \Gamma_i$ in $\mathcal{E}$.*
2. *q has all pushforwards and fiberwise coproducts, i.e., for any operation $g : \Gamma \to A$ and list of colors $\Omega \sqsubset \Gamma$ there is a pushforward $g\,\Omega \sqsubset A$, and for any family of colors $(R_i \sqsubset A)_{i\in I}$, there is a fiberwise coproduct $\sum_{i\in I} R_i \sqsubset A$.*

Moreover, the equivalence holds while maintaining any bound $|I| < \kappa$ on the cardinalities of the indexing sets.

[8] Cf. [MZ 2013, p. 13], [Shulman 2023, Thm. 4.28]

Polynomial closure

We say q is **polynomially closed** when either of the equivalent conditions holds with $\kappa = \omega$, meaning colors of $\mathcal{E}$ are closed under taking finite sums of monomials "weighted" by operations of $\mathcal{B}$.

Proposition

The following identities hold

$$\sum_{i \in I} \sum_{j \in J} R_{ij} \equiv \sum_{(i,j) \in I \times J} R_{ij}$$

$$f(\Theta, \sum_{i \in I} R_i, \Theta') \equiv \sum_{i \in I} f(\Theta, R_i, \Theta')$$

$$f(g_1 \, \Omega_1, \ldots, g_n \, \Omega_n) \equiv (f \circ (g_1, \ldots, g_n))(\Omega_1, \ldots, \Omega_n)$$

in the sense that whenever the minimal lift on one side exists then so does the other, with a canonical isomorphism between them.

Polynomial closure

Let Set be the operad of sets and n-ary functions.

Let Subset be the operad whose colors are pairs $(X, U \subset X)$, and whose operations $(X_1, U_1), \ldots, (X_n, U_n) \to (Y, V)$ are functions $f : X_1, \ldots, X_n \to Y$ st $x_1 \in U_1, \ldots, x_n \in U_n \Rightarrow f(x_1, \ldots, x_n) \in V$. Let $\mathrm{sub} : \mathrm{Subset} \to \mathrm{Set}$ be the evident projection.

Proposition

sub *is polynomially closed, where pushforward and fiberwise coproducts are given by image and union respectively:*

$$f\left((X_1, U_1), \ldots, (X_n, U_n)\right) = (Y, f(U_1, \ldots, U_n))$$

$$\sum_{i \in I} (X, V_i) = (X, \cup_{i \in I} V_i)$$

Model of a gCFG

Let $p : \mathsf{F}\,\mathcal{S} \to \mathcal{O}$ be a functor of operads, w/associated map of species $\phi : \mathcal{S} \to \mathcal{O}$. Let $q : \mathcal{E} \to \mathcal{B}$ be any functor of operads. A **model** of p in q is a commuting square

$$\begin{array}{ccc} \mathsf{F}\,\mathcal{S} & \xrightarrow{\quad \tilde{M} \quad} & \mathcal{E} \\ {\scriptstyle p}\big\downarrow & & \big\downarrow{\scriptstyle q} \\ \mathcal{O} & \xrightarrow{\quad M \quad} & \mathcal{B} \end{array}$$

such that for every color R of $\mathsf{F}\,\mathcal{S}$, the cone of nodes in $\mathcal{S}$

$$(x : R_1, \dots, R_k \overset{\phi}{\underset{g}{\Longrightarrow}} R)_{x \in \mathcal{S}}$$

is mapped to a q-minimal cone in $\mathcal{E}$

$$(\tilde{M}_x : \tilde{M}_{R_1}, \dots, \tilde{M}_{R_k} \overset{q}{\underset{M_g}{\Longrightarrow}} \tilde{M}_R)_{x \in \mathcal{S}}$$

Model of a gCFG

A model of a gCFG G is a model of its underlying functor p.

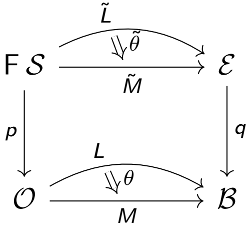
Thus, in our sum-of-pushforward notation, a model $(M, \tilde{M})$ of a gCFG corresponds to a solution for the system of equations

$$\tilde{M}_R \equiv \sum_{R_1,\ldots,R_k \Rightarrow_g^{\phi} R} M_g(\tilde{M}_{R_1}, \ldots, \tilde{M}_{R_k}) \tag{3}$$

with one such equation for every non-terminal.

The category of models

Let $(L, \tilde{L})$ and $(M, \tilde{M})$ be models of p in q. A **morphism** $(L, \tilde{L}) \Rightarrow (M, \tilde{M})$ is given by a pair of natural transformations $\theta : L \Rightarrow M$ and $\tilde{\theta} : \tilde{L} \Rightarrow \tilde{M}$ such that the diagram commutes



in the sense that the natural transformations obtained by whiskering are equal $q \circ \tilde{\theta} = \theta \circ p$.

The category of models

Note the definition does not impose any compatibility conditions between the natural transformations $(\theta, \tilde{\theta})$ and the minimal cones in q, in other words it is just a **2-morphism**

$$(\theta, \tilde{\theta}) \quad : \quad (L, \tilde{L}) \Longrightarrow (M, \tilde{M}) \quad : \quad p \to q$$

between the underlying morphisms of functors.

Given arbitrary functors p and q, we write $[p, q]$ for the category of morphisms of functors $p \to q$ and 2-morphisms between them. When $p : \mathsf{F}\,\mathcal{S} \to \mathcal{O}$ is a functor from a free operad, we write $\mathrm{Mod}(p, q)$ for the full subcategory of $[p, q]$ spanned by the models.

The language generated by a gCFG as an initial model

We aim to show that the language of constants generated by a gCFG G defines an initial model of G in sub : $\mathrm{Subset} \to \mathrm{Set}$.

We will obtain this as a corollary of several more basic facts, and in particular via a more fundamental ("proof-relevant") model of G in the polynomially closed functor tgt : $\mathrm{Set}^{\to} \to \mathrm{Set}$.

The constants algebra

Every operad $\mathcal{O}$ comes equipped with a canonical functor

$$\mathrm{el}[\mathcal{O}] : \mathcal{O} \to \mathrm{Set}$$

(abbreviated "el" when $\mathcal{O}$ is clear from context), defined by

$$\begin{aligned} \mathrm{el}_A &= \{\, c \mid c : A \,\} \\ \mathrm{el}_f &= (c_1, \ldots, c_n) \mapsto f \circ (c_1, \ldots, c_n) \end{aligned}$$

For example when $\mathcal{O} = \mathsf{W}\,\mathcal{C}$:

$$\begin{aligned} \mathrm{el}_{(A,B)} &= \mathcal{C}(A, B) \\ \mathrm{el}_{w_0 - \ldots - w_n} &: \mathcal{C}(A_1, B_1) \times \cdots \times \mathcal{C}(A_n, B_n) \to \mathcal{C}(A, B) \\ \mathrm{el}_{w_0 - \ldots - w_n} &= (u_1, \ldots, u_n) \mapsto w_0 u_1 w_1 \ldots u_n w_n \end{aligned}$$

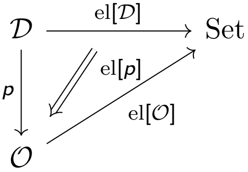
The constants algebra

A functor $\mathcal{O} \to \mathrm{Set}$ is also called a $\mathcal{O}$-algebra.

Important fact: el is the *initial $\mathcal{O}$-algebra*, in the sense that it has a unique natural transformation to any other algebra $M : \mathcal{O} \to \mathrm{Set}$, defined by the family of fns $\mathrm{el}_A \to M_A$ sending a constant $c : A$ of $\mathcal{O}$ to the element M_c of M_A determined by the algebra structure.

The constants algebra

For any functor $p : \mathcal{D} \to \mathcal{O}$, we can therefore build a triangle



where $\mathrm{el}[p]$ is uniquely determined by initiality of $\mathrm{el}[\mathcal{D}]$.

Arrow operads

In general, natural transformations $\theta : L \Rightarrow M : \mathcal{O} \to \mathcal{P}$ between functors of operads have the following equivalent description.

Let $\mathcal{P}^{\to}$ be the operad whose colors are unary operations u of $\mathcal{P}$, and whose n-ary operations $u_1, \ldots, u_n \to u$ are pairs (f_s, f_t) of n-ary operations of $\mathcal{P}$ such that $f_t \circ (u_1, \ldots, u_n) = u \circ f_s$.

There are two evident functors $\mathrm{src}, \mathrm{tgt} : \mathcal{P}^{\to} \to \mathcal{P}$.

Then giving a natural transformation $\theta : L \Rightarrow M : \mathcal{O} \to \mathcal{P}$ is equivalent to giving a functor of operads $\tilde{\theta} : \mathcal{O} \to \mathcal{P}^{\to}$ such that $\mathrm{src} \circ \tilde{\theta} = L$ and $\mathrm{tgt} \circ \tilde{\theta} = M$.

An initial model in $\mathrm{tgt} : \mathrm{Set}^{\to} \to \mathrm{Set}$

The canonical natural transformation $\mathrm{el}[p] : \mathrm{el}[\mathcal{D}] \Rightarrow \mathrm{el}[\mathcal{O}] \circ p$ therefore induces a commutative square:

$$\begin{array}{ccc} \mathcal{D} & \xrightarrow{\widetilde{\mathrm{el}}[p]} & \mathrm{Set}^{\to} \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle \mathsf{tgt}} \\ \mathcal{O} & \xrightarrow{\mathrm{el}[\mathcal{O}]} & \mathrm{Set} \end{array}$$

Theorem: this defines an initial model of p in tgt!

Polynomial closure of $\mathtt{tgt}$

Proposition

$\mathtt{tgt}$ *is polynomially closed, where:*

$$\begin{aligned} f\,(u_1 : Y_1 \to X_1, \ldots, u_n : Y_n \to X_n) &= f \circ (u_1, \ldots, u_n) \\ &\quad : Y_1 \times \cdots \times Y_n \to X \\ \sum_{i \in I} (v_i : Y_i \to X) &= [v_i]_{i \in I} : \coprod_{i \in I} Y_i \to X \end{aligned}$$

Initiality of the constants model

Two key facts:

1. For any $p : \mathcal{D} \to \mathcal{O}$, the morphism $(\mathrm{el}[\mathcal{O}], \widetilde{\mathrm{el}}[p]) : p \to \mathsf{tgt}$ is an initial object in $[p, \mathsf{tgt}]$.
2. $(\mathrm{el}[\mathcal{O}], \widetilde{\mathrm{el}}[p])$ is a model of p in tgt when $\mathcal{D} = \mathsf{F}\,\mathcal{S}$.

(1) is immediate. (2) relies on inductive definition of $\mathsf{F}\,\mathcal{S}$.

Since $\mathrm{Mod}(p, \mathsf{tgt})$ is a full subcategory of $[p, \mathsf{tgt}]$, we conclude that $(\mathrm{el}[\mathcal{O}], \widetilde{\mathrm{el}}[p])$ is an initial object in $\mathrm{Mod}(p, \mathsf{tgt})$!

An initial model in $\mathsf{sub} : \mathrm{Subset} \to \mathrm{Set}$

Consider the composite morphism:

$$\begin{array}{ccccc}
\mathsf{F}\,\mathcal{S} & \xrightarrow{\widetilde{\mathrm{el}}[p]} & \mathrm{Set}^{\to} & \xrightarrow{\mathrm{im}} & \mathrm{Subset} \\
{\scriptstyle p}\downarrow & & \downarrow{\scriptstyle \mathsf{tgt}} & & \downarrow{\scriptstyle \mathsf{sub}} \\
\mathcal{O} & \xrightarrow{\mathrm{el}[\mathcal{O}]} & \mathrm{Set} & = & \mathrm{Set}
\end{array}$$

This defines an initial model of p in sub, essentially because the image functor is a left adjoint.[9]

We recover the "language of constants" as $L_G = \mathrm{im}(\widetilde{\mathrm{el}}_p(S))$!

[9] Postcomposition with the left adjoint morphism $\mathrm{im} : \mathsf{tgt} \to \mathsf{sub}$ induces a functor $[p, \mathrm{im}] : [p, \mathsf{tgt}] \to [p, \mathsf{sub}]$ which is itself a left adjoint, and therefore sends the initial object $(\mathrm{el}_{\mathcal{O}}, \widetilde{\mathrm{el}}_p)$ to an initial object $(\mathrm{el}_{\mathcal{O}}, \widetilde{\mathrm{el}}_p\,\mathrm{im})$ in $[p, \mathsf{sub}]$. Moreover, it may be readily verified that im preserves pushforwards and fiberwise coproducts, and hence preserves models. (More generally, a left adjoint morphism into a functor of operads with all minimal lifts preserves minimal cones.) We conclude that $(\mathrm{el}_{\mathcal{O}}, \widetilde{\mathrm{el}}_p\,\mathrm{im})$ is an initial model of p in sub.

A more abstract view of gCFLs

The initial model of a grammar G in $\text{tgt} : \mathrm{Set}^{\to} \to \mathrm{Set}$ may be seen as a "proof-relevant language", in the sense that it encodes not just a subset of constants generated by G but also the set of parse trees of every constant in the language.

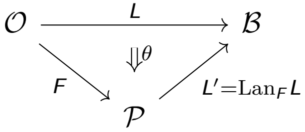
But why stop there? Given any functor of operads $q : \mathcal{E} \to \mathcal{B}$ we can *define* the language generated by G in q, notated $\tilde{L}^q_G$, as the interpretation $\tilde{L}_S \sqsubset^q L_A$ of its start symbol $S \sqsubset^p A$ for some initial model $(L, \tilde{L}) : p \to q$.

We refer to the languages generated by gCFGs in q as q-**gCFLs**.

Some closure properties of q-gCFLs

If $q : \mathcal{E} \to \mathcal{B}$ is polynomially closed, then q-gCFLs are closed under (the appropriate analogue of) "union" and "combination", defined using fiberwise coproduct and pushforward respectively.[10]

If $\mathcal{B}$ is moreover cocomplete, then q-gCFLs are closed under "functorial image", defined using pushforward along a natural transformation obtained via left Kan extension.



[10] To make these statements precise, we need to be able to refer to the *base interpretation* $L : \mathcal{O} \to \mathcal{B}$ of a q-gCFL $(L, \tilde{L}) : p \to q$. For example, "union" is stated as follows: if $\tilde{L}_1, \ldots, \tilde{L}_k \sqsubset^q L_A$ are q-gCFLs with the same base interpretation L, then $\sum_{i=1}^k \tilde{L}_i \sqsubset^q L_A$ is a q-gCFL with base interpretation L.

Conclusion

Summary

Several steps of generalization and abstraction:

$$\text{CFGs} \quad \rightsquigarrow \quad \text{CFGs of arrows} \begin{array}{c} \mathsf{F}\,\mathcal{S} \\ \Big\downarrow \\ \mathsf{W}\,\mathcal{C} \end{array} \quad \rightsquigarrow \quad \text{gCFGs} \begin{array}{c} \mathsf{F}\,\mathcal{S} \\ \Big\downarrow \\ \mathcal{O} \end{array}$$

$$\text{CFLs} \quad \rightsquigarrow \quad \text{initial models of CFGs} \quad \rightsquigarrow \quad q\text{-gCFLs}$$

Some open questions and directions

1. Other interesting examples of gCFGs?
2. Interesting examples of *q*-gCFLs for *q* other than tgt or sub?
3. Are *q*-gCFLs closed under intersection with *q*-regular languages? (Surely yes, but we need the right definitions!)
4. What are pushdown automata in this setting?[11]
5. When does a gCFG have a unique model?[12]
6. Is there a nice story to tell about SOL definability?

[11] Ongoing work with PAM, which we need to resume!

[12] Some results with F. Jafarrahmani, which we need to write up!

Extra slides

Decomposing the intersection of a gCFL with a regular language

Given a gCFG and a NDFA over the same operad, we obtain a pullback in Operad from a corresponding pullback in $\mathrm{Species}$:

$$\begin{array}{ccc} \mathsf{F}\,\mathcal{S}' & \xrightarrow{\mathsf{F}\,\psi} & \mathsf{F}\,\mathcal{S} \\ {\scriptstyle p'}\downarrow & \lrcorner & \downarrow{\scriptstyle p} \\ \mathcal{Q} & \xrightarrow[p_{\mathcal{Q}}]{} & \mathcal{O} \end{array} \qquad \Bigg| \qquad \begin{array}{ccc} \mathcal{S}' & \xrightarrow{\psi} & \mathcal{S} \\ {\scriptstyle \phi'}\downarrow & \lrcorner & \downarrow{\scriptstyle \phi} \\ \mathcal{Q} & \xrightarrow[p_{\mathcal{Q}}]{} & \mathcal{O} \end{array}$$

This relies crucially on the fact that $p_{\mathcal{Q}}$ is finitary and ULF!

Taking image of gCFL generated by p' along $p_{\mathcal{Q}}$ yields intersection.