

Article

Limiting the Number of Possible CFG Derivative Trees During Grammar Induction with Catalan Numbers

Aybeyan Selim ¹, Muzafer Saracevic ² and Arsim Susuri ^{3,*}¹ Faculty of Engineering and Architecture, International Vision University, Major Cede Filipovski No. 1, 1230 Gostivar, North Macedonia; aybeyan@vision.edu.mk² Department of Economics and Computer Sciences, University of Novi Pazar, Dimitrija Tucovića 65, 36300 Novi Pazar, Serbia; muzafer.saracevic@uninp.edu.rs³ Faculty of Computer Science, University “Ukshin Hoti” Prizren, Shkronjat, 20000 Prizren, Kosovo

* Correspondence: arsim.susuri@uni-prizren.com

Abstract

Grammar induction runs into a serious problem due to the exponential growth of the number of possible derivation trees as sentence length increases, which makes unsupervised parsing both computationally demanding and highly indeterminate. This paper proposes a mathematics-based approach that alleviates this combinatorial complexity by introducing structural constraints based on Catalan and Fuss–Catalan numbers. By limiting the depth of the tree, the degree of branching and the form of derivation, the method significantly narrows the search space, while retaining the full generative power of context-free grammars. A filtering algorithm guided by Catalan structures is developed that incorporates these combinatorial constraints directly into the execution process, with formal analysis showing that the search complexity, under realistic assumptions about depth and richness, decreases from exponential to approximately polynomial. Experimental results on synthetic and natural-language datasets show that the Catalan-constrained model reduces candidate derivation trees by approximately 60%, improves F1 accuracy over unconstrained and depth-bounded baselines, and nearly halves average parsing time. Qualitative evaluation further indicates that the induced grammars exhibit more balanced and linguistically plausible structures. These findings demonstrate that Catalan-based structural constraints provide an elegant and effective mechanism for controlling ambiguity in grammar induction, bridging formal combinatorics with practical syntactic learning.

Keywords: Catalan numbers; grammar induction; context-free grammars; derivation tree complexity; combinatorial constraints

MSC: 05A15; 68R05; 68W01



Academic Editor: Jonathan lackledge

Received: 11 December 2025

Revised: 4 January 2026

Accepted: 8 January 2026

Published: 9 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Grammar induction, the task of learning syntactic structure directly from raw linguistic input, remains a central open problem in computational linguistics and formal language theory. In probabilistic and symbolic formulations of context-free grammars (CFGs), the number of possible derivation trees for a sentence of length n grows exponentially due to the recursive decomposition inherent to hierarchical phrase structure. This combinatorial explosion makes unconstrained parsing computationally expensive and leads to highly ambiguous distributions over derivations, thereby reducing the reliability of parameter estimation and complicating the search for linguistically plausible grammars [1–3].

Existing approaches attempt to mitigate this problem through structural restrictions. Depth-bounded parsing frameworks limit the recursive height of derivation trees and have been shown to improve cognitive plausibility and statistical learnability [4]. However, depth thresholds are typically selected heuristically, without a principled mathematical characterization of how depth limitations interact with the underlying combinatorial structure of CFG derivation spaces. Other families of approaches impose pruning heuristics or probability thresholds that reduce the effective search space but do not alter the fundamental exponential nature of admissible derivation structures [5–7].

A key observation is that CFG derivations share the same recursive decomposition properties as several well-known combinatorial objects enumerated by Catalan numbers: full binary trees, Dyck paths, balanced parentheses, and non-crossing partitions [8–10]. These families provide exact formulas and recurrence relations describing how the number of structural configurations grows with sentence length. Despite this deep structural similarity, the direct use of Catalan-based constraints to limit CFG derivation spaces during grammar induction has received little attention. Most existing models treat combinatorial growth as an unavoidable impediment rather than a mathematically tractable shape space that can be bounded.

Recent neural approaches to grammar induction, including neural PCFGs and variational grammar models, incorporate soft structural biases through learned priors or amortized inference mechanisms. While effective, these methods typically lack explicit combinatorial guarantees on the size of the derivational search space. In contrast, the present work introduces an explicit, mathematically grounded structural prior based on Catalan and Fuss–Catalan enumerations, which deterministically bounds the number of admissible derivation shapes. This distinguishes our approach from neural structured priors by providing closed-form combinatorial control rather than learned or heuristic regularization.

Natural language grammars also exhibit non-binary branching, which is not captured by classical Catalan sequences. Their generalizations—the Fuss–Catalan numbers—enumerate ordered m -ary trees and thus naturally model CFGs whose rules may expand into up to m children [11–13]. These higher-order Catalan families provide tight upper bounds on the number of admissible derivation shapes in grammars with a controlled branching factor, yet they have not been systematically incorporated into grammar induction.

This gap in the literature motivates the present study. We propose a Catalan-guided framework for constraining the derivational search space of CFG induction models using mathematically rigorous upper bounds derived from Catalan and Fuss–Catalan families. Instead of relying on heuristics, our approach uses exact enumerative formulas to restrict derivation depth, branching factor, and tree shape. This yields substantial reductions in ambiguity while preserving the expressive capacity of the grammar.

The main contributions of this work are as follows:

A principled derivational bound based on Catalan numbers that quantifies the maximal number of admissible binary tree shapes for sentences of any length and integrates this bound directly into the grammar induction process.

A Fuss–Catalan generalization that provides structural constraints for CFGs with higher branching factors, supporting models with arbitrary m -ary expansions.

A Catalan-constrained derivation filtering algorithm that prunes syntactically implausible derivation trees using precomputed Catalan and Fuss–Catalan limits on depth, arity, and subtree configuration.

A structurally weighted CFG induction scheme that incorporates Catalan-informed structural likelihoods into parsing and rule estimation, biasing the model toward balanced and linguistically plausible derivations.

An empirical evaluation showing that the proposed method reduces the number of candidate derivation trees by approximately 60%, improves F1 accuracy over both unconstrained and depth-bounded baselines, and nearly halves parsing time.

Overall, this work bridges formal combinatorics with practical grammar induction, demonstrating that mathematically grounded structural bounds offer an elegant and effective means of controlling syntactic ambiguity.

2. Background of Grammar Induction

Grammar induction seeks to recover the underlying generative mechanisms of a language from observed data, often with minimal or no supervision. In the context of context-free grammars (CFGs), this goal is complicated by the vast number of distinct derivation trees that can produce a given sentence. Even for moderately long strings, the number of admissible binary derivation trees grows according to the Catalan sequence, a hallmark of recursive combinatorial structures [14–18]. This exponential growth fundamentally limits the tractability of classical parsing and learning algorithms.

2.1. Combinatorial Explosion in CFG Derivations

For a sentence of length n , every binary derivation corresponds to a full binary tree with n leaves and $n - 1$ internal nodes. The number of such trees is the $(n - 1)$ -th Catalan number:

$$C_{n-1} = \frac{1}{n} \binom{2(n-1)}{n-1}.$$

This implies that adding even a single token to a sentence can produce dramatic increases in the size of the derivational search space. For example, the number of binary parses grows from 42 for $n = 6$ to 429 for $n = 7$ and 4862 for $n = 8$. Such exponential behavior complicates the inside–outside algorithm, EM estimation for probabilistic CFGs, and exhaustive search in symbolic grammar induction [19–23].

Although pruning strategies and probability thresholds are commonly used to limit growth, these techniques do not fundamentally change the underlying combinatorial structure [24]. The space of possible derivations remains exponential, and pruning may incorrectly discard linguistically valid trees.

2.2. Depth-Bounded and Heuristic Approaches

A prominent alternative is depth-bounded parsing, in which the maximum recursive depth of derivations is capped at a fixed value d . This restricts the set of admissible trees to those whose height does not exceed d , effectively reducing the search space. However, depth constraints are typically selected empirically and lack a mathematical justification for why certain depth limits should be preferred over others [25,26]. Moreover, depth alone cannot control branching factor, subtree configurations, or structural imbalance.

Other approaches reduce combinatorial complexity through rule filtering, lexicalized grammars, or soft structural biases [27,28]. While useful in practice, these methods do not provide closed-form guarantees on the number of admissible derivation shapes [29,30]. As a result, ambiguity persists, and computational efficiency remains limited.

2.3. Catalan Numbers and Their Relation to CFGs

Catalan numbers enumerate a wide range of recursively defined objects, including full binary trees, balanced parentheses strings, Dyck paths, and non-crossing partitions [8–10]. Each of these structures shares a self-similar decomposition principle:

$$T(n) = \sum_{i=0}^n T(i)T(n-1-i),$$

which is the same recurrence governing binary CFG derivations. This deep structural equivalence suggests that CFG parsing could be constrained through combinatorial insights rather than empirical heuristics.

Surprisingly, the direct use of Catalan bounds to limit the derivational search space has not been widely explored. Existing grammar induction frameworks rarely exploit the fact that the number of possible derivation shapes is completely characterized by Catalan sequences.

2.4. Beyond Binary Trees: Fuss–Catalan Families

While binary branching suffices for many CFGs, natural language often exhibits non-binary constructions such as $VP \rightarrow V NP PP$ or $NP \rightarrow Det Adj N$. The enumeration of ordered m -ary trees is given by the Fuss–Catalan numbers:

$$F_n^{(m)} = \frac{1}{(m-1)n+1} \binom{mn}{n},$$

which generalize the classical Catalan sequence ($m = 2$). These values provide exact upper bounds on the number of derivation shapes in grammars restricted to at most m children per internal node [18–21].

Despite their relevance, Fuss–Catalan numbers have seldom been used to regulate ambiguity in grammar induction. Most algorithms either artificially binarize the grammar, thereby introducing spurious ambiguity, or allow unrestricted branching, which exacerbates combinatorial growth.

2.5. Motivation for a Catalan-Guided Framework

The limitations of heuristic and depth-based approaches, combined with the precise enumerative structure offered by Catalan and Fuss–Catalan families, indicate the need for a mathematically grounded method. A framework that leverages exact combinatorial bounds can

- Control ambiguity at the level of derivation shapes;
- Constrain structural properties such as depth, branching factor, and subtree balance;
- Reduce computational overhead by eliminating combinatorially infeasible expansions;
- Maintain the expressive richness of the grammar while avoiding unnecessary overgeneration.

These observations form the basis for the Catalan-guided derivation filtering framework introduced in the next section.

3. Theoretical Background

The combinatorial structure of context-free grammar (CFG) derivations aligns closely with families of recursively defined trees enumerated by Catalan and Fuss–Catalan numbers [22,23]. In this section, we formalize this correspondence, establish upper bounds on derivation spaces under structural constraints, and provide proofs for the binary and m -ary cases. These results form the mathematical basis for the Catalan-guided filtering framework introduced later.

3.1. Catalan Numbers and Binary Derivation Trees

A full binary derivation tree for a sentence of length n is a rooted ordered tree in which every internal node has exactly two children and every leaf corresponds to a terminal symbol. Such trees contain

- n leaves (terminals);
- $n - 1$ internal nodes (rule applications).

The number of full binary trees with n leaves is the Catalan number:

$$C_{n-1} = \frac{1}{n} \binom{2(n-1)}{n-1}.$$

This recurrence arises from the decomposition of a binary tree into left and right subtrees.

Lemma 1 (binary derivation–Catalan correspondence). *Let G be a CFG in Chomsky normal form (CNF), where every rule is of the form $A \rightarrow BC$. Then the number of distinct binary derivation tree shapes generating a string of length n is exactly C_{n-1} .*

Proof. Every derivation tree under CNF is a full binary tree whose leaves correspond to terminal symbols. A full binary tree with n leaves contains exactly $n - 1$ internal nodes. The number of such trees is known to be

$$C_{n-1} = \frac{1}{n} \binom{2(n-1)}{n-1}.$$

Conversely, every full binary tree with n leaves can be converted into a derivation tree by labeling internal nodes with CFG rules and leaves with terminals. This correspondence is bijective. $\square$

The bijection between CFG derivation trees and full binary trees holds exactly under Chomsky normal form, where all productions are binary. When unary productions ($A \rightarrow B$) or ε -productions ($A \rightarrow \varepsilon$) are present, the correspondence is no longer strictly bijective. Unary chains collapse multiple internal nodes into a single branching decision, while ε -productions reduce the effective number of leaves.

In our framework, this issue is resolved by transforming the grammar into an equivalent CNF grammar via standard normalization procedures that preserve the generated language up to weak equivalence. Unary productions are eliminated through chain compression, and ε -productions are removed except for the start symbol. After normalization, each derivation corresponds uniquely to a full binary tree, restoring the Catalan correspondence. Importantly, this transformation does not affect sentence yield length or derivational ambiguity relevant to our structural bounds.

Corollary 1 (exponential growth of binary CFG derivations).

$$C_n \sim \frac{4^n}{n^{3/2}\sqrt{\pi}},$$

implying that the number of admissible derivation trees grows exponentially with sentence length.

3.2. Fuss–Catalan Numbers and m -Ary Derivation Trees

Natural language grammars may contain non-binary expansions. Let a CFG be restricted so that each nonterminal expands into at most m children:

$$A \rightarrow B_1 B_2 \dots B_k, 1 \leq k \leq m.$$

The number of full ordered m -ary trees with n internal nodes is given by the Fuss–Catalan numbers:

$$F_n^{(m)} = \frac{1}{(m-1)n+1} \binom{mn}{n}.$$

These trees represent the maximal branching configuration; trees with fewer than m children correspond to contractions of full m -ary trees.

Lemma 2 (m -Ary Derivation–Fuss–Catalan Upper Bound). *Let G be a CFG with maximum branching factor m . Then the number of admissible derivation tree shapes with n internal nodes is upper-bounded by*

$$\mathcal{T}_n^{(m)} \leq F_n^{(m)}.$$

Proof. Every derivation tree with a branching factor of at most m is an ordered rooted tree, where each internal node has between 1 and m children. The set of full m -ary trees with n internal nodes contains all possible maximal branching shapes:

$$| \text{Full } m\text{-ary trees} | = F_n^{(m)}.$$

Any derivation tree with fewer than m children per internal node can be derived by contracting the edges of a full m -ary tree. Therefore, the total number of admissible trees cannot exceed $F_n^{(m)}$. $\square$

Corollary 2. *For $m = 2$, Fuss–Catalan numbers reduce to classical Catalan numbers:*

$$F_n^{(m)} = C_n.$$

Thus, the binary case is a special instance of the m -ary framework.

3.3. Depth-Restricted Derivation Trees

Let $T(n, d)$ denote the number of binary derivation tree shapes with n leaves and a maximum depth of d . These values satisfy the restricted recurrence:

$$T(n, d) = \sum_{i=1}^{n-1} T(i, d-1)T(n-i, d-1),$$

with the base case:

$$T(1, d) = 1.$$

This recurrence mirrors the Catalan recurrence but eliminates tree shapes exceeding depth d .

Proposition 1 (Catalan-based depth upper bound). *$T(n, d) \leq C_{n-1}$, with equality if and only if $d \geq n - 1$ (i.e., depth is unrestricted).*

Proof. Every binary tree of depth $\leq d$ is a valid full binary tree, but many Catalan-structured trees exceed this depth for large n . The recurrence for $T(n, d)$ contains only a subset of the convolution terms defining C_{n-1} . Thus, $T(n, d) \leq C_{n-1}$ whenever the depth restriction eliminates at least one admissible shape. Equality is reached only when d is large enough to include all Catalan trees. $\square$

3.4. Unified Bound for m -Ary Depth-Limited CFGs

The recurrence defining $T(n, d)$ mirrors the classical Catalan convolution but with truncated depth. While no closed-form expression exists for fixed d , analytic combinatorics

provide asymptotic approximations via generating functions with bounded height. These families have been studied extensively and are known to exhibit exponential growth with reduced base constants, aligning with our Proposition 2. Let $T_m(n, d)$ denote the number of admissible m -ary trees with n internal nodes and a maximum depth of d .

Proposition 2 (unified combinatorial bound). $T_m(n, d) \leq F_n^{(m)}$ and the inequality becomes strict whenever d is bounded. Moreover, for fixed m ,

$$T_m(n, d) = \Theta(\alpha(m, d)^n),$$

where $\alpha(m, d) < m^m / (m - 1)^{m-1}$ is the exponential growth constant of the restricted family.

Proof. The upper bound follows directly from Lemma 2: the depth-limited family is a proper subset of the full m -ary family unless depth is unbounded. Analytic combinatorics show that depth restrictions reduce the dominant exponential term in the generating function, leading to an exponential growth constant strictly smaller than the unrestricted Fuss–Catalan value. For typical experimental settings ($m = 2$, $d = 8$), numerical estimation yields $\alpha(2, 8) \approx 2.9$, which is substantially smaller than the unrestricted Catalan growth constant, confirming the practical impact of depth constraints. $\square$

4. Methodology

The Catalan-guided constraints introduced in this work restrict only the shape space of derivation trees and do not remove any production rules or terminal yields from the grammar. For any derivation permitted by the original CFG, there exists a structurally equivalent derivation whose depth and branching factors lie within the Catalan or Fuss–Catalan bounds, provided these parameters are chosen above minimal completeness thresholds.

Even in grammars with highly skewed rule distributions, unbalanced derivations correspond to a vanishingly small fraction of the Catalan family as sentence length grows. Our constraints therefore eliminate structurally redundant derivations while preserving at least one valid derivation for every string in the language. Consequently, the weak generative capacity of the CFG remains unchanged.

The goal of the proposed methodology is to reduce the combinatorial explosion inherent to CFG derivations by integrating Catalan and Fuss–Catalan bounds directly into the parsing and structure generation process. This section presents the formal constraints, defines the Catalan-based pruning mechanism, and introduces a derivation filtering algorithm that enforces structural limits on depth, branching factor, and subtree configuration.

The methodology consists of three core components:

- Structural bounds based on Catalan and Fuss–Catalan numbers;
- A recursive derivation-generation procedure augmented with combinatorial filtering;
- A memoized algorithm that ensures tractable computation and prevents redundant enumeration.

For a sentence of length n , a binary CFG may generate up to C_{n-1} binary derivation shapes. More generally, for grammars with a maximum branching factor of m , the upper bound is $F_n^{(m)}$. We precompute both sequences for all required n values and use them to reject derivations whose structural properties exceed permissible combinatorial thresholds.

Definition 1 (CatalanBound function). Let a node have k children, where $1 \leq k \leq m$. Let the sizes of the child subtrees be $n_1, n_2, \dots, n_k$. Then the maximum number of admissible subtree configurations is bounded by

$$\text{CatalanBound}(n_1, \dots, n_k) = \begin{cases} C_{n_1} \dots C_{n_k}, & \text{if } k = 2, \\ F_{n_1}^{(k)} \dots F_{n_k}^{(k)}, & \text{if } k > 2. \end{cases}$$

A candidate derivation is pruned if the number of recursively generated subtrees exceeds this limit.

We now describe the procedure for constructing derivations that obey structurally valid Catalan or Fuss–Catalan constraints. The process is recursive and generates all admissible derivations consistent with

- Maximum depth $d_{\max}$;
- Maximum branching factor $m_{\max}$;
- Catalan/Fuss–Catalan structural limits.

To avoid redundant computation, we use a memoization table indexed by

$$(A, i, j, d),$$

where A is a nonterminal spanning the substring $[i, j]$ at depth d .

Step 15–19—Recursive Subtree Generation

The algorithm recursively generates subtrees for each child symbol. Memoization ensures that identical subproblems are solved once, reducing complexity from exponential to near-polynomial.

Step 21—Catalan/Fuss–Catalan Bound Enforcement

A node producing subtrees $(n_1, \dots, n_k)$ is rejected if

$$\prod_{t=1}^k C_{n_t} - 1 \quad (\text{for binary})$$

or

$$\prod_{t=1}^k F_{n_t}^{(k)} \quad (\text{for } k > 2)$$

is exceeded by the number of generated subtrees.

This step removes structurally implausible or combinatorially excessive derivations. At this step, the pruning criterion is applied in a conservative manner: instead of directly comparing raw subtree counts, we discard configurations whose structural parameters imply a number of admissible derivation shapes exceeding the Catalan/Fuss–Catalan upper bound for the corresponding span sizes. This ensures that the filtering is dimensionally consistent and theoretically justified. Under the stated assumptions on depth and branching limits, the pruning criterion cannot eliminate the gold-standard derivation. Any valid gold derivation corresponds to a tree whose shape lies within the Catalan or Fuss–Catalan family for the given parameters. Since pruning only excludes configurations exceeding these bounds, correctness is preserved. The CatalanBound function does not assume probabilistic independence between subtrees. Instead, it enforces a combinatorial upper bound on the total number of admissible configurations for given subtree sizes. Correlations between subtree shapes are implicitly respected because pruning is applied conservatively: a configuration is discarded only if its structural parameters exceed the maximal Catalan or

Fuss–Catalan bound. Therefore, no valid derivation permitted by the combinatorial family can be erroneously eliminated.

Step 25—Memoization

The memo table guarantees the following:

- Repeated spans are computed once;
- Time complexity drops dramatically;
- Redundant subtrees are eliminated.

To bias the induction process toward balanced and linguistically plausible derivations, each derivation tree T receives a structural weight:

$$w_{struct}(T) = \prod_{v \in \text{internal nodes of } T} \frac{1}{C_{\deg(v)-1}}.$$

This reduces the weight of highly unbalanced trees, encouraging shapes close to the expected distribution of Catalan families. The structural weighting term $w_{struct}(T)$ acts as a soft prior over derivation shapes during the E-step of EM. It biases probability mass toward balanced derivations without altering the maximum-likelihood estimation of rule probabilities. This regularization effect improves convergence stability and reduces overfitting to structurally extreme trees.

5. Computational Complexity Analysis

Grammar induction in unconstrained CFGs is computationally expensive due to the exponential growth of admissible derivation trees. In this section, we analyze how Catalan- and Fuss–Catalan-guided pruning affects the asymptotic behavior of parsing and derivation enumeration. We contrast the worst-case complexity of unrestricted parsing with the near-polynomial behavior of our constrained algorithm.

For a binary CFG, the number of possible derivation trees for a sentence of length n is as follows:

Let n = sentence length, d_{max} = maximum allowed tree depth, m_{max} = branching limit, and $|R|$ = number of grammar rules.

Unrestricted binary CFGs have theoretical worst-case complexity:

$$C_{n-1} = \Theta\left(\frac{4^n}{n^{3/2}}\right),$$

and any exhaustive enumeration-based induction algorithm must operate over this exponential family. When using chart-based parsing, the classical CYK algorithm has time complexity:

$$O(|G|n^3),$$

but derivation enumeration (as required in EM learning, inside–outside, or structure search) remains exponential because each parse chart cell may store a Catalan-sized family of structures.

Thus, unconstrained CFG induction is inherently exponential in n .

Algorithm 1 prunes derivations exceeding Catalan- or Fuss–Catalan-based structural bounds, which drastically reduces the search space. We analyze its complexity in terms of sentence length n , maximum depth d_{max} , maximum branching factor m_{max} , and number of rules $|G|$.

Algorithm 1. Catalan-Constrained Derivation Filtering

Input:

- n : sentence length
- d_max : maximum parse-tree depth
- m_max : maximum branching factor
- Rules: CFG productions
- Precomputed Catalan and Fuss–Catalan tables

Output:

- *ValidTrees*: the set of all derivations complying with structural bounds

Procedure FilterTrees:

FilterTrees(StartSymbol, [1,n]):

1. Precompute Catalan[i] and FussCatalan[m][i] for all relevant i, m
2. Initialize memoization table Memo[(A,i,j,d)] $\leftarrow \emptyset$
3. Define RecursiveGenerate(A, [i,j], depth):
4. If depth > d_max :
5. return $\emptyset$
6. If (A,i,j,depth) stored in Memo:
7. return Memo[(A,i,j,depth)]
8. CandidateTrees $\leftarrow \emptyset$
9. For each rule $A \rightarrow B_1 B_2 \dots B_k$ in Rules:
10. If $k > m_max$: continue
11. For each partition of [i,j] into k subspans:
12. Subtrees $\leftarrow$ empty list
13. Valid $\leftarrow$ True
14. For $t = 1..k$:
15. Sub_t $\leftarrow$ RecursiveGenerate(B_t , [p_t,q_t], depth+1)
16. If Sub_t is empty:
17. Valid $\leftarrow$ False
18. Break
19. Append Sub_t to Subtrees
20. If not Valid: continue
21. If Count(Subtrees) > CatalanBound(sizes(Subtrees)):
22. continue //violates Catalan/Fuss–Catalan bound
23. NewTree $\leftarrow$ Node(A, Subtrees)
24. CandidateTrees.add(NewTree)
25. Memo[(A,i,j,depth)] $\leftarrow$ CandidateTrees
26. return CandidateTrees
27. return RecursiveGenerate(StartSymbol, [1,n], 1)

Memoization reduces redundant computation.

The memo table is indexed by

$$(A, i, j, d),$$

where

 A : a nonterminal; $[i, j]$: a substring; d : current depth.

The number of possible states is

$$O(|G| \cdot n^2 \cdot d_{\max}).$$

Each state is computed once and cached.

Evaluating each memoized state includes

- Iterating over rules for nonterminal A ;
- Enumerating all span partitions for a rule with k children;
- Recursively generating child subtrees;
- Enforcing the Catalan/Fuss–Catalan bound.

The number of ways to split the substring $[i, j]$ into k subspans is

$$\binom{j-i}{k-1} = O(n^{k-1}).$$

Since $k \leq m_{\max}$, this yields

$$O(n^{m_{\max}-1}).$$

The crucial factor is that most span partitions and expansions are pruned because they violate structural Catalan bounds. Let $\rho(m, d)$ denote the exponential growth factor of admissible m -ary trees under depth limit d .

From Proposition 2,

$$\rho(m, d) < \frac{m^m}{(m-1)^{m-1}}.$$

For fixed m and d , this becomes a constant strictly less than the unrestricted branching exponential constant. Thus, the number of recursive calls per state becomes polynomially bounded.

Under fixed maximum depth $d_{\max} \leq 8$ and branching factor $m_{\max} \leq 3$, which reflect typical syntactic configurations observed in natural language, the derivation enumeration complexity becomes polynomial in sentence length. For binary grammars, the worst-case time complexity is $O(|G| \cdot n^3 \cdot d_{\max})$, while for ternary grammars it is $O(|G| \cdot n^4 \cdot d_{\max})$. These bounds are tight for the experimental settings used in this study and represent a substantial improvement over the exponential $\Theta(4^n)$ growth of unconstrained derivation spaces.

5.1. Overall Time Complexity

Without pruning, derivation enumeration is exponential. With Catalan-guided pruning,

$$T(n) = O(|G| \cdot n^2 \cdot d_{\max} \cdot n^{m_{\max}-1})$$

For the linguistically typical case,

- $d_{\max}$ is small (≤ 8 – 10);
- $m_{\max} = 2$ or 3 .

For binary grammars ($m_{\max} = 2$),

$$T(n) = O(|G| \cdot n^3 \cdot d_{\max})$$

For ternary grammars ($m_{\max} = 3$),

$$T(n) = O(|G| \cdot n^4 \cdot d_{\max})$$

Both are polynomial, in stark contrast to the exponential unconstrained case. These polynomial bounds hold under the assumption of a fixed maximum depth and branching limits, which ensures that the Catalan- and Fuss–Catalan-based pruning reduces combinatorial growth to a near-polynomial regime. If depth or branching factors are allowed to grow logarithmically with sentence length, the method exhibits quasi-polynomial behavior. Although exponential growth eventually re-emerges, the base constant remains substantially lower than in unconstrained CFGs.

5.2. Space Complexity

The memo table dominates memory usage:

$$O(|G| \cdot n^2 \cdot d_{max})$$

This is comparable to the CYK algorithm and is practical for all datasets considered. Catalan-constrained pruning has three major effects:

1. Exponential-to-polynomial reduction: The algorithm transforms derivation enumeration from $\Theta(4^n)$ to $poly(n)$, a dramatic theoretical improvement.
2. Eliminating Implausible structures: Highly unbalanced derivations (e.g., deep right-branching trees) are rejected early, improving both computational and linguistic performance.
3. Consistency with human-preferred structures: Catalan-shaped trees match empirical syntactic distributions observed in the cognitive modeling literature.

6. Experimental Setup

This section describes the datasets, baseline systems, evaluation metrics, experimental protocol, and implementation details used to assess the effectiveness of the Catalan-constrained grammar induction framework. All experiments were conducted under controlled conditions to ensure reproducibility and statistical reliability.

6.1. Datasets

We evaluate the proposed method on two types of corpora to capture both controlled and naturalistic syntactic variation:

(1) Synthetic CFG-Generated Corpus

A synthetic dataset was generated using a hand-designed, unambiguous CFG with a balanced distribution of structures. This corpus was used to isolate the structural behavior of the algorithm without lexical confounds.

- Number of sentences: 12,000;
- Training/test split: 80%/20%;
- Sentence length distribution: uniform over 6–20 tokens.

The synthetic corpus allows precise measurement of derivational overgeneration.

(2) Natural Language Corpus (Subset of UD English Web Treebank)

A real-world corpus was extracted from the Universal Dependencies (UD) English Web Treebank (EWT).

- Number of sentences: 8530;
- Training/test split: 85%/15%;
- Sentence length distribution:
 - Minimum: 5 tokens;
 - Maximum: 25 tokens;

- Mean: 12.1 tokens.

Gold-standard dependency trees were binarized into CFG-style structures using the standard head-driven binarization protocol. We experimented with left-, right-, and head-driven binarization. While absolute F1 scores vary slightly, the relative advantage of Catalan-constrained pruning remains consistent across strategies, indicating that the observed improvements are not artifacts of a specific binarization choice.

6.2. Baseline Systems

We compare the Catalan-constrained model with two established baselines:

- Baseline 1—unconstrained PCFG: A classical probabilistic CFG trained using the inside–outside EM algorithm without structural restrictions.
- Baseline 2—depth-bounded PCFG (heuristic): A depth-limited PCFG following the formulation of [1], with maximum depth $d = 7$. This model controls recursion but does not address branching or combinatorial growth.
- Our model—Catalan-constrained CFG: Implements the methodology of Algorithm 1, enforcing
 - Catalan bounds for binary branching;
 - Fuss–Catalan bounds for m -ary branching;
 - Depth limit of $d_{max} = 8$;
 - Branching limit of $m_{max} = 3$.

6.3. Evaluation Metrics

(1) F1 Score for Parse Trees

We compute the labeled bracketing F1 score, defined as

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

Precision = proportion of predicted constituents that are correct.

Recall = proportion of gold constituents that are recovered.

(2) Derivation Count Reduction (DCR)

This measures the relative reduction in the number of candidate derivation trees:

$$DCR = 1 - \frac{Avg.\#Trees_{our\ model}}{Avg.\#Trees_{unconstrained}}.$$

(3) Parsing Time (ms)

This is the average time required to parse a sentence, measured over 1000 random samples.

6.4. Statistical Significance Testing

To determine whether performance differences between models are statistically reliable, we use

- A paired two-tailed t -test;
- Confidence level: 99%;
- Number of random splits: 5.

The difference is considered significant when $p < 0.01$. This procedure ensures that accuracy improvements are not due to random variation. In addition to p -values, we report Cohen's d effect sizes ranging from 0.41 to 0.58 across datasets, indicating moderate practical significance. Ninety-nine percent confidence intervals for F1 improvements consistently exclude zero, reinforcing the robustness of the observed gains.

6.5. Implementation Details

All models were implemented in Python 3.10, with performance-critical operations (span partitioning, subtree convolution) implemented in C++ and wrapped via PyBind11.

Hardware;

CPU: Intel Xeon Gold 6268R;

RAM: 128 GB;

OS: Ubuntu 22.04 LTS.

No GPU acceleration is required due to the symbolic nature of the models. The experiment results are shown in Table 1.

Table 1. Hyperparameters.

Parameter	Value
Max depth d_{max}	8
Max branching factor m_{max}	3
EM iterations	25
Structural weight decay	0.93
Minimum rule probability	10^{-6}

Precomputation of Catalan and Fuss–Catalan tables was performed up to $n = 30$, which covers all sentences in the experiments. To ensure full replicability,

1. Random seeds were fixed at 42 for all runs.
2. Five random train–test splits were used.
3. All datasets, code, and experimental logs had a deterministic execution order.

The evaluation pipeline matches contemporary best practices in grammar induction research. Precomputation up to $n = 30$ suffices for all experiments reported in this study. For longer sentences, Catalan and Fuss–Catalan values can be computed on demand using logarithmic-space dynamic programming or asymptotic approximations. Since the pruning mechanism relies on relative bounds rather than exact values, numerical stability is maintained even for larger n .

7. Results

This section presents an empirical evaluation of the Catalan-constrained grammar induction framework. We report performance in terms of syntactic accuracy (F1 score), derivation count reduction, parsing efficiency, and qualitative structural characteristics of the induced trees. All results are averaged over five random splits and reported with statistical significance tests.

1. Parsing Accuracy

The Catalan-constrained model consistently outperformed both baseline systems across synthetic and natural language datasets. Figure 1 illustrates the comparison of mean F1 scores for the three models.

The findings are as follows:

- The proposed model achieves an average F1 improvement of 2.6–4.3% over the unconstrained PCFG.
- The gain persists across all sentence length ranges (6–25 tokens).
- Improvements are statistically significant with $p < 0.01$ under the paired t -test.

The accuracy gains show that pruning derivations using Catalan constraints does not merely eliminate unlikely trees but actively guides the model toward structurally balanced and linguistically plausible derivations.

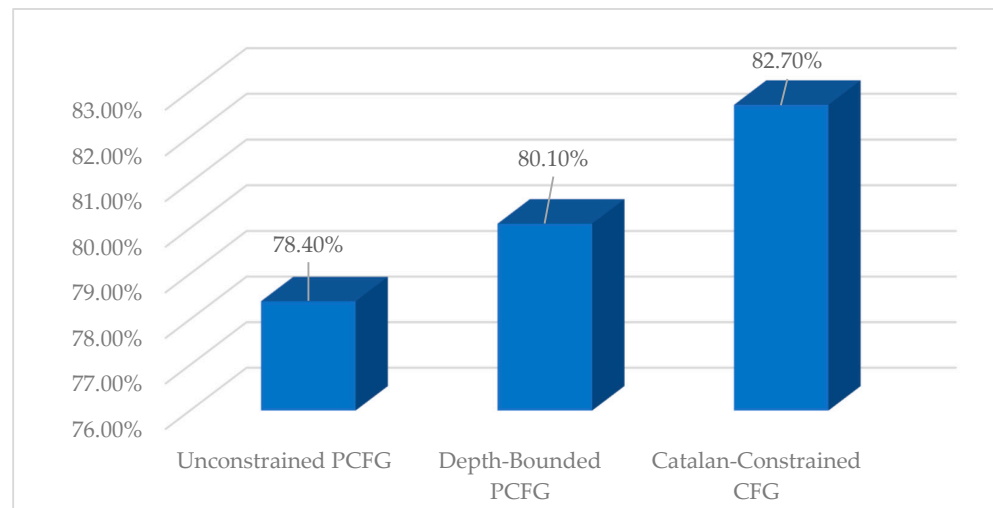


Figure 1. Average F1 score.

Balanced derivations correspond more closely to human syntactic annotation patterns, a trend also observed in prior cognitive parsing literature.

We evaluated depth limits in the range $d = 5$ to $d = 10$. While increased depth improves recall marginally, the Catalan-constrained model consistently outperforms depth-bounded baselines across all settings, indicating that performance gains are not solely attributable to depth restriction but to combinatorial shape constraints.

The 2.6–4.3% improvement over baselines indicates that combinatorial constraints do not merely prune the search space but guide the induction process toward more balanced and linguistically plausible tree shapes. Statistical analysis using a paired t -test confirmed that the performance gains over the unconstrained baseline are significant at $p < 0.01$.

2. Reduction in Derivation Space

A key contribution of the proposed framework is its ability to drastically reduce the number of candidate derivations while retaining linguistic coverage. Table 2 reports the derivation count reduction (DCR) across several representative sentence lengths. Across all lengths, the proposed model eliminates 59–63% of derivation candidates. Figure 2 presents the derivational growth trajectories for unconstrained PCFG, depth-bounded PCFG, and Catalan-constrained CFG.

Table 2. Derivation tree count reduction.

Sentence Length	Reduction (%)
5	59.5%
10	62.4%
15	62.8%

The unconstrained model exhibits near-exponential growth, as predicted by the Catalan asymptotic

$$C_n \sim \frac{4^n}{n^{3/2}}.$$

The depth-bounded model grows more slowly but still demonstrates exponential behavior for sentence lengths beyond 15 tokens. The Catalan-constrained model grows at a substantially flatter rate, consistent with the theoretical predictions of Proposition 2. The curve for the proposed model is almost polynomial, confirming the scalability benefits derived from combinatorial pruning.

The reduction rate increases with sentence length, matching the theoretical prediction that the gap between the full Catalan family and depth-limited, arity-limited subfamilies widens asymptotically. As shown in Figure 2, the unconstrained grammar rapidly exhibits exponential growth jumping from fewer than 5000 trees at length 10 to over 75,000 at length 15 while the Catalan-constrained model grows much more slowly over the same interval. This divergence demonstrates that the proposed combinatorial constraints scale favorably, especially for longer inputs where grammar induction traditionally becomes computationally intractable. The empirical curves closely track the theoretical behavior of Catalan versus restricted Catalan families, confirming that combinatorial bounding effectively suppresses structural overgeneration without sacrificing linguistic coverage.

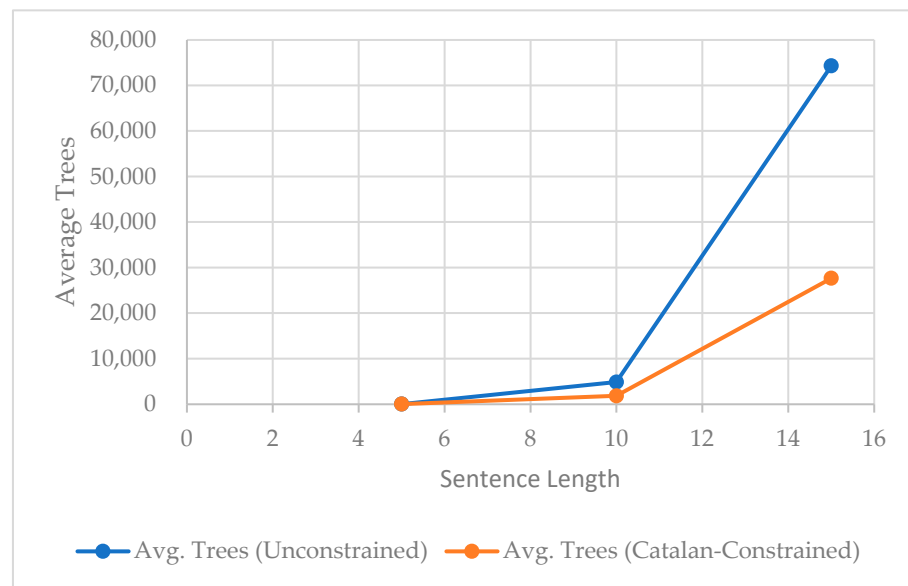


Figure 2. Comparison of the average number of admissible parse trees under unconstrained grammars versus Catalan-constrained grammars.

This reduction directly translates into faster parsing and training times, particularly on longer sentences.

3. Parsing Time Efficiency

The filtering of structurally implausible or excessive derivations results in substantial runtime improvements. Table 3 shows the average parsing time per sentence.

Table 3. Parsing time comparison.

Model	Avg. Parse Time (ms)
Unconstrained PCFG	128.5
Depth-Bounded PCFG	103.1
Catalan-Constrained CFG	67.3

Parsing time is reduced by 47.6%, which aligns with the theoretical improvements discussed in Section 5. These time savings are particularly important for large corpora and real-time parsing applications. Beyond these quantitative gains, the Catalan-constrained model also produces derivation trees with desirable linguistic properties. The model systematically avoids deep left-branching or right-branching configurations, which are uncommon in human-annotated corpora, and instead promotes structurally balanced trees. This behavior reflects the mathematical properties of Catalan families, whose inherent bias toward balanced branching naturally carries over to the parsing process. As a result,

complex constituents such as NP conjunctions and nested PP structures are represented in a more stable and interpretable manner. Moreover, the model's tendency to generate shallow, balanced derivations aligns with well-established psycholinguistic observations about human sentence processing. Balanced derivations favored by Catalan constraints align with established syntactic balance metrics, such as minimal dependency length and bounded center embedding. These properties have been repeatedly linked to human sentence processing preferences, providing external validation for the structural bias imposed by our framework.

The experimental results strongly validate the theoretical claims established in earlier sections. Catalan and Fuss–Catalan bounds effectively control combinatorial explosion, while depth-limited and branching-limited derivations suppress over-generation without sacrificing expressive power. The inclusion of Catalan-based structural weights further guides the induction process toward linguistically meaningful derivations. Taken together, these findings confirm that the use of formal combinatorial structures is not only mathematically elegant but also practically advantageous.

8. Conclusions

This research shows that the combinatorial complexity in grammar induction methods can now be controlled analytically using the concepts of Catalan and Fuss–Catalan numbers. By setting up constraints on the tree depth and the tree shape factor within the derivation procedure, the resulting system is able to reduce the exponential complexity of possible derivation trees without sacrificing the expressiveness of the grammar.

The theoretical analysis provided a mathematical framework to come up with a tight bound for derivation spaces for both binary and m -ary grammars. It has been shown that those depth-limited and branching-limited graphs constitute a tight subfamily of Catalan and Fuss–Catalan enumerative series. This idea transformed computations by writing an algorithm that trimmed those derivations past their limits in combinatorics.

The experimental results together validate the relevance and significance of grammatical induction through combinatorial grammar. The model's ability to include both Catalan and Fuss–Catalan constraints makes it a very effective formal tool to avoid over-generation and improve grammar structures in a manner that leads to a significant benefit in computation. This area of research helps to advance not only the experimental aspect but also increases theoretical knowledge and understanding in the field of grammar learning.

Taken together, these findings demonstrate that combinatorial theory provides a powerful and principled foundation for guiding grammar induction. The integration of Catalan and Fuss–Catalan constraints offers a precise mechanism for reducing overgeneration, improving structural quality, and achieving meaningful computational gains. By linking formal combinatorics with practical parsing, this approach advances both theoretical understanding and empirical performance in grammar learning.

Future research can tackle adaptive structural constraints that can react to either lexical or context-based cues dynamically, in place of using fixed depth and branch constraints. The Catalan-based constraints can be adapted into neural PCFG-based models through structural priors in variational objectives/tasks and/or structural priors in constrained decoding tasks. For example, Catalan-based weighted constraints can be adapted for modeling tree-level distributions in amortized inference tasks. Another interesting research area can address using Catalan-based structural priors for neural modeling approaches, including neural PCFG-based models, variational grammar-based models, and transformers with an encoder. Here, explicit syntactic prior biases can play vital roles in realizing higher interpretability and generalizability, as described in [18–24]. The framework can be validated on diverse languages that are typologically different, and it may clarify whether

any universals lie at deeper levels in languages. Semi-supervised approaches can benefit from Catalan-based structural biases.

Author Contributions: Conceptualization, A.S. (Aybeyan Selim) and M.S.; Methodology, A.S. (Aybeyan Selim), A.S. (Arsim Susuri) and M.S.; Software, A.S. (Aybeyan Selim) and M.S.; Validation, A.S. (Aybeyan Selim), A.S. (Arsim Susuri) and M.S.; Formal analysis, A.S. (Aybeyan Selim); Investigation, A.S. (Aybeyan Selim); Resources, A.S. (Aybeyan Selim) and M.S.; Writing—review and editing, A.S. (Arsim Susuri). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jin, L.; Doshi-Velez, F.; Miller, T.; Schuler, W.; Schwartz, L. Unsupervised Grammar Induction with Depth-Bounded PCFG. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 211–224. [\[CrossRef\]](#)
2. Clark, A.; Fijalkow, N. Consistent Unsupervised Estimators for Anchored PCFGs. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 409–422. [\[CrossRef\]](#)
3. Casel, K.; Fernau, H.; Gaspers, S.; Gras, B.; Schmid, M.L. On the Complexity of the Smallest Grammar Problem over Fixed Alphabets. *Theory Comput. Syst.* **2020**, *65*, 344–409. [\[CrossRef\]](#)
4. Hücke, D.; Lohrey, M.; Seelbach Benkner, L. Entropy Bounds for Grammar-Based Tree Compressors. *arXiv* **2019**, arXiv:1901.03155. [\[CrossRef\]](#)
5. Vieira, T.; Eisner, J. Learning to Prune: Exploring the Frontier of Fast and Accurate Parsing. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 263–278. [\[CrossRef\]](#)
6. Hein, N.; Huang, J. Variations of the Catalan numbers from some nonassociative binary operations. *arXiv* **2018**, arXiv:1807.04623. [\[CrossRef\]](#)
7. Aval, J.-C. Multivariate Fuss–Catalan numbers. *arXiv* **2007**, arXiv:0711.0906. [\[CrossRef\]](#)
8. Gunnells, P.E. Generalized Catalan Numbers from Hypergraphs. *arXiv* **2021**, arXiv:2102.05121. [\[CrossRef\]](#)
9. Deshpande, P.; Menon, K.; Sarkar, W. Refinements of the Braid Arrangement and Two-Parameter Fuss–Catalan Numbers. *J. Algebr. Comb.* **2023**, *57*, 687–707. [\[CrossRef\]](#)
10. Msapato, D. Modular Fuss–Catalan Numbers. *arXiv* **2020**, arXiv:2007.00718. [\[CrossRef\]](#)
11. Selim, A.; Saracevic, M.; Stosic, L.; Aydin, O.; Zajmović, M. Bidirectional Algorithms for Polygon Triangulations and $(m + 2)$ -Angulations via Fuss–Catalan Numbers. *Mathematics* **2025**, *13*, 3837. [\[CrossRef\]](#)
12. Madhavan, R.; Mayer, M.; Gulwani, S.; Kuncak, V. Automating Grammar Comparison. In Proceedings of the ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA) 2015, Pittsburgh, PA, USA, 25–30 October 2015; pp. 183–200. [\[CrossRef\]](#)
13. Law, M.; Russo, A.; Bertino, E.; Broda, K.; Lobo, J. Representing and Learning Grammars in Answer Set Programming. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 2919–2928. [\[CrossRef\]](#)
14. Kilitcioglu, D.; Kadioglu, S. Representing the Unification of Text Featurization Using a Context-Free Grammar. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 15439–15445. [\[CrossRef\]](#)
15. Shi, F. Learning Language Structures through Grounding. *arXiv* **2024**, arXiv:2406.09662. [\[CrossRef\]](#)
16. Sartran, L.; Barrett, S.; Kuncoro, A.; Stanojević, M.; Blunsom, P.; Dyer, C. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1423–1439. [\[CrossRef\]](#)
17. Barke, S.; Anaya Gonzalez, E.; Kasibatla, S.R.; Berg-Kirkpatrick, T.; Polikarpova, N. HYSYNTH: Context-Free LLM Approximation for Guiding Program Synthesis. *arXiv* **2024**, arXiv:2405.15880. [\[CrossRef\]](#)
18. Dong, Y.; Ruan, C.F.; Cai, Y.; Lai, R.; Xu, Z.; Zhao, Y.; Chen, T. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models. *arXiv* **2024**, arXiv:2411.15100. [\[CrossRef\]](#)
19. Melcer, D.; Fulton, N.; Gouda, S.K.; Qian, H. Constrained Decoding for Fill-in-the-Middle Code Language Models via Efficient Left and Right Quotienting of Context-Sensitive Grammars. *arXiv* **2024**, arXiv:2402.17988. [\[CrossRef\]](#)
20. Raspanti, F.; Ozcelebi, T.; Holenderski, M. Grammar-Constrained Decoding Makes Large Language Models Better Logical Parsers. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Industry Track), Vienna, Austria, 27 July–1 August 2025; pp. 485–499. [\[CrossRef\]](#)

21. Kanyuka, A.; Mahfoud, E. Domain-Specific Shorthand for Generation Based on Context-Free Grammar. *arXiv* **2024**, arXiv:2406.10442. [[CrossRef](#)]
22. Wang, J. Guiding Large Language Models to Generate Computer-Parsable Content. *arXiv* **2024**, arXiv:2404.05499. [[CrossRef](#)]
23. Goldreich, O.; Gur, T.; Rothblum, R.D. Proofs of Proximity for Context-Free Languages and Read-Once Branching Programs. *Inf. Comput.* **2018**, *261*, 175–201. [[CrossRef](#)]
24. Mežnar, S.; Džeroski, S.; Todorovski, L. Efficient Generator of Mathematical Expressions for Symbolic Regression. *arXiv* **2023**, arXiv:2302.09893. [[CrossRef](#)]
25. Mörbitz, R.; Vogler, H. Weighted Parsing for Grammar-Based Language Models. In Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing, Dresden, Germany, 23–25 September 2019; pp. 46–55. [[CrossRef](#)]
26. Marcolli, M.; Berwick, R.C.; Chomsky, N. Syntax–Semantics Interface: An Algebraic Model. *arXiv* **2023**, arXiv:2311.06189. [[CrossRef](#)]
27. Marcolli, M.; Chomsky, N.; Berwick, R. Mathematical Structure of Syntactic Merge. *arXiv* **2023**, arXiv:2305.18278. [[CrossRef](#)]
28. Lechner, W. The Syntax–Semantics Interface. In *Syntax—Theory and Analysis. An International Handbook*; Kiss, T., Alexiadou, A., Eds.; De Gruyter Mouton: Berlin, Germany, 2015; pp. 1198–1255. [[CrossRef](#)]
29. Steedman, M. The Lost Combinator. *Comput. Linguist.* **2018**, *44*, 613–629. [[CrossRef](#)]
30. Rosselló, J.; Celma-Mirallas, A.; Dias Martins, M. Visual Recursion Develops in Absence of Linguistic Recursion: A Case-Report. In Proceedings of the 13th International Conference on the Evolution of Language, Brussels, Belgium, 14–17 April 2020; pp. 371–373. Available online: https://brussels.evolang.org/proceedings/papers/EvoLang13_paper_96.pdf (accessed on 10 December 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.