Syntax Error Correction as Idempotent Matrix Completion

ANONYMOUS AUTHOR(S)

In this work, we illustrate how to lower context-free language recognition onto a tensor algebra over finite fields. In addition to its theoretical value, this connection has yielded surprisingly useful applications in incremental parsing, code completion and program repair. To accelerate code completion, we design and implement a novel incremental parser-synthesizer that transforms CFGs onto a dynamical system over finite field arithmetic, enabling us to suggest syntax repairs in-between keystrokes. Our approach attains state-of-the-art repair precision@$\{5, 10\}$ on a dataset of human Python repairs at a fraction of the cost of existing methods. Unlike prior work, our approach does not employ neural networks, and is therefore highly flexible to additional constraints and readily extensible to new programming languages.

1 INTRODUCTION

In programming, most errors initially manifest as syntax errors, and though often cosmetic, manual repair can present a significant challenge for novice programmers. *Syntax correction* is the problem of repairing a syntactically incorrect program so that it parses. This problem is challenging because well-formed programs have many extra-syntactic constraints, increasing the complexity of repair, but also because the problem is highly under-determined: repairs must be consistent with the surrounding context, and the correct repair is seldom unique, even assuming minimality.

The majority of prior work on syntax correction can be separated into one of two high-level categories: (1) *pattern matching*, which uses hand-crafted rules to identify and correct common syntax errors, and (2) *machine learning*, i.e., which uses a statistical language model to repair code via a set of learned heuristics. The former approach is effective but requires language-specific rules. The latter generates more natural edits, but is costly to train, generalizes poorly, has high sample complexity, and is difficult to incorporate new constraints without retraining.

Traditionally, parsers have ignored developer tools like code completion, to their detriment. The approach we propose offers a compelling alternative to its textbook presentation in the parsing literature, not only because it unifies parsing, code completion and error correction under a simple algebraic framework, but also because it is composable using ordinary logical primitives, which are highly flexible to additional constraints and well-suited for SAT-based implementation.

In our work, we recast the problem of syntax correction as a special case of tensor completion with an logical semiring, and lay the foundation for a new approach to program repair grounded in formal language theory, unifying *code completion*, *error correction* and *syntax repair*. We provide exact and approximate algorithms for solving these problems in the real-world setting, and implement them in a real-time editor called Tidyparse, demonstrating their practical utility. Given a well-formed grammar, our tool can be used to complete unfinished code, parse incomplete code and repair broken fragments in arbitrary context-free and linear conjunctive languages.

More specifically, our paper is structured as follows: we present two high-level approaches to syntax correction: one that samples random edits and accepts only those which parse (Theory 1, probabilistic correction), and another that uses equational reasoning to find the satisfying assignments to a system of multilinear equations over finite fields (Theory 2, model-theoretic). Finally, we show how these two approaches can be combined to attain state-of-the-art performance (Theory 1.5), and validate it on a variety of real-world program repair scenarios.

POPL, January 17-19, 2024, London, United Kingdom
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

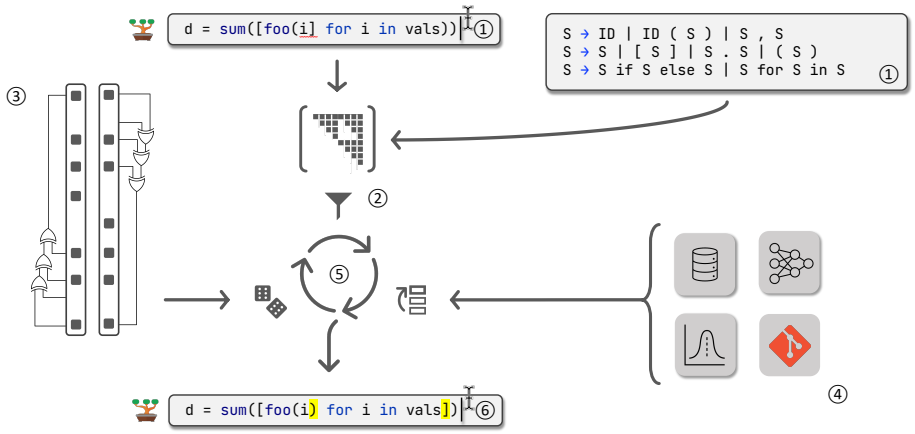


Fig. 1. Our framework consists of several components, which work together to repair strings. An invalid string and grammar (1) are first compiled into a Boolean tensor (2), representing a linear system of equations whose fixed points characterize the admisisble set. The system can be solved directly, yielding a set of strings which are then ranked using a suitable scoring function (4). Optionally, we may introduce stochastic edits to the string using the Levenshtein ball sampler (3) and decode the admissible set incrementally by resampling (5). When the langauge edit distance and solutions are time-sensitive, (5) is typically more efficient.

2 OVERVIEW

We now present a number of optimality criteria for the parsing framework. These criteria were used to guide the design of various components and evaluate its performance.

- First and foremost, the framework must be **sound** and **complete**. That is, the parser-generator must (1) accept arbitrary conjunctive grammars, and (2) generate a parser which accepts all and only syntactically valid strings and (3) when a string is invalid, our synthesizer must eventually generate every syntactically valid string within a fixed distance from it.
- Second, we require that the resulting repairs be **plausible** and **diverse**. In other words, the framework should generate repairs that are likely to be written by a human being, consistent with the surrounding context, and reasonably diverse in their structure.
- Third, the framework must be **efficient** and **responsive**. That is, it must be able to recognize well-formed strings in subcubic time, and generate admissible repairs in subpolynomial time. These conditions are necessary to provide real-time feedback whilst the user is typing.
- Fourth, the framework must also be **robust** and **scalable**. In practice, this means that the framework should be robust to multiple errors, handle grammars with a large number of productions and be able to scale linearly as the number of processor cores are increased.
- Finally, the framework must be **flexible** and **extensible**. Intended as a general-purpose tool for generating syntax repairs in a wide variety of programming languages, end-users should be able to extend the framework with their own custom grammars and side-constraints.

2.1 Natural language explanation

Our architecture accepts a grammar and a string, which may contain any arbitrary characters. We tokenize the string into nonterminals and attempt to parse using Valiant's algorithm [30]. If the string is syntactically valid, we return the parse forest, otherwise, we return the partial parse forest

and a small set of edits that would make the string syntactically valid. This set is always guaranteed to exist, by overwriting the invalid string with the shortest string in the language.

If the string is invalid, we first form a bijection between edits and the Levenshtein ball using a combinatorial number system, sample without replacement using an LFSR with leapfrog partitioning to distribute the work across cores, and decode the resulting bitvectors to enumerate edit templates. This sequence can be viewed as a permutation of a space-filling curve over the space of strings within a fixed edit distance of the original, invalid, string.

We then use each edit template, in conjunction with the grammar to search for admissible repairs. This is done by encoding the template and grammar as a matrix equivalence relation and solving for the least fixpoint using a linear system of equations over a finite field. We describe this process in more detail in Section 4. The resulting samples will all be syntactically valid according to the grammar. These are reranked according to a statistical distance metric, and we output the most likely samples to the user. This whole process is illustrated in Figure 1.

3 PROBLEM STATEMENT

The problem of syntax error correction under a finite number of typographic errors is reducible to the bounded Levenshtein CFL reachability problem, which can be formally stated as follows:

Definition 3.1. The language edit distance (LED) is the minimum number of edits required to transform an invalid string into a valid one, where validity is defined as containment in a context-free language, $\ell : \mathcal{L}$, i.e., $\Delta^*(\underset{\sim}{\sigma}, \ell) \coloneqq \min_{\sigma \in \ell} \Delta(\underset{\sim}{\sigma}, \sigma)$, and Δ is the Levenshtein distance.

We seek to find the set of strings S such that $\forall \tilde{\sigma} \in S, \Delta(\underset{\sim}{\sigma}, \tilde{\sigma}) \leq k$, where k is the maximum number of edits greater than or equal to the language edit distance. We call this set the *Levenshtein ball* of $\underset{\sim}{\sigma}$ and denote it $\Delta_k(\underset{\sim}{\sigma})$. Since $1 \leq \Delta^*(\underset{\sim}{\sigma}, \ell) \leq k$, we have $1 \leq k$. We now consider an upper bound on $\Delta^*(\underset{\sim}{\sigma}, \ell)$, i.e., the greatest lower bound on k such that $\Delta_k(\underset{\sim}{\sigma}) \cap \ell \neq \varnothing$.

Lemma 3.2. *For any nonempty language $\ell : \mathcal{L}(\mathcal{G})$ and invalid string $\underset{\sim}{\sigma} : \Sigma^n$, there exists an $(\tilde{\sigma}, m)$ such that $\tilde{\sigma} \in \ell \cap \Sigma^m$ and $0 < \Delta(\underset{\sim}{\sigma}, \ell) \leq \max(m, n) < \infty$.*

Proof. Since ℓ is nonempty, it must have at least one inhabitant $\sigma \in \ell$. Let $\tilde{\sigma}$ be the smallest such member. Since $\tilde{\sigma}$ is a valid sentence in ℓ, by definition it must be that $|\tilde{\sigma}| < \infty$. Let $m \coloneqq |\tilde{\sigma}|$. Since we know $\underset{\sim}{\sigma} \notin \ell$, it follows that $0 < \Delta(\underset{\sim}{\sigma}, \ell)$. Let us consider two cases, either $\tilde{\sigma} = \varepsilon$, or $0 < |\tilde{\sigma}|$:

- If $\tilde{\sigma} = \varepsilon$, then $\Delta(\underset{\sim}{\sigma}, \tilde{\sigma}) = n$ by full erasure of $\underset{\sim}{\sigma}$, or
- If $0 < m$, then $\Delta(\underset{\sim}{\sigma}, \tilde{\sigma}) \leq \max(m, n)$ by overwriting.

In either case, it follows $\Delta(\underset{\sim}{\sigma}, \ell) \leq \max(m, n)$ and ℓ is always reachable via a finite nonempty set of Levenshtein edits, i.e., $0 < \Delta(\underset{\sim}{\sigma}, \ell) < \infty$. □

Let us now consider the maximum growth rate of the admissible set, $A \coloneqq \Delta_k(\underset{\sim}{\sigma}) \cap \ell$, as a function of k and n. Let $\ell \coloneqq \underset{\sim}{\sigma}$. Since ℓ is finite and thus regular, $\bar{\ell} = \Delta_k(\underset{\sim}{\sigma}) \setminus \{\underset{\sim}{\sigma}\}$ is regular by closure of regular languages under complementation, and thus context-free a fortiori. Since $\bar{\ell}$ accepts every string except $\underset{\sim}{\sigma}$, it represents the worst CFL in terms of asymptotic growth of A.

Lemma 3.3. *The complexity of searching A is upper bounded by $\mathcal{O}\left(\sum_{c=1}^{k} \binom{cn+n+1}{c} (|\Sigma| + 1)^c\right)$.*

Proof. We can overestimate the size of A by considering the number of unique ways to insert, delete, or substitute c terminals into a string $|\underset{\sim}{\sigma}|$ of size n. This can be overaproximated by interleaving ε^c around every token, i.e., $\underset{\sim}{\sigma}' \coloneqq (\varepsilon^c \underset{\sim}{\sigma}_i)_{i=1}^{n} \varepsilon^c$, where $|\underset{\sim}{\sigma}'| = cn + n + 1$, and only considering substitution. We augment $\Sigma_\varepsilon \coloneqq \Sigma \cup \{\varepsilon\}$ so that deletions and insertions may be considered as special cases of substitution. Thus, we have $cn + n + 1$ positions to substitute ($|\Sigma_\varepsilon|$) tokens, i.e.,

$\binom{cn+n+1}{c}(|\Sigma_\varepsilon|)^c$ ways to edit σ' for each $c \in [1,k]$. This upper bound is not tight, as overcounts many identical edits w.r.t. σ. Nonetheless, it is sufficient to show $|A| < \sum_{c=1}^{k} \binom{cn+n+1}{c}(|\Sigma_\varepsilon|)^c$. □

We note that the above bound applies to all strings and languages, and relates to the Hamming bound on $H_k(\sigma)$, which only considers substitutions. In practice, much tighter bounds may be obtained by considering the structure of ℓ and σ. For example, based on an empirical evaluation from a dataset of human errors and repairs in Python code snippets ($|\Sigma| = 50, |\sigma| < 40, \Delta(\sigma, \ell) \in [1,3]$), we estimate the *filtration ratio*, i.e., the density of the admissible set relative to the Levenshtein ball, $D = |A|/|\Delta_k(\sigma)|$ to have empirical mean $E_\sigma[D] \approx 2.6 \times 10^{-4}$, and variance $\mathrm{Var}_\sigma[D] \approx 3.8 \times 10^{-7}$.

In practice, this problem is ill-posed even when $k = \Delta^*(\sigma, \ell) \approx 1$. For example, consider the language of ursine dietary preferences. Although $\sigma :=$ "Bears like to eat tomatoes" is not in the language, e.g., $\tilde{\sigma} :=$"Bears like to eat" is valid ($\Delta^* = 1$), however there are many other variations with the same or similar edit distance, e.g., or "Bears like to eat {berries, honey, fish}", or "{Polar, Panda} bears like to eat {seals, bamboo}". In general, there are many valid strings within a fixed Levenshtein distance of σ, and we seek to find among them, the set of strings that are both syntactically valid and semantically plausible within a short amount of time.

4 MATRIX THEORY

Recall that a CFG is a quadruple consisting of terminals (Σ), nonterminals (V), productions ($P \colon V \to (V \mid \Sigma)^*$), and a start symbol, ($S$). It is a well-known fact that every CFG is reducible to *Chomsky Normal Form*, $P' \colon V \to (V^2 \mid \Sigma)$, in which every production takes one of two forms, either $w \to xz$, or $w \to t$, where $w, x, z : V$ and $t : \Sigma$. For example:

$$\mathcal{G} := \{\, S \to S\,S \mid (\,S\,) \mid (\,) \,\} \Longrightarrow \mathcal{G}' = \{\, S \to Q\,R \mid S\,S \mid L\,R, \quad R \to), \quad L \to (, \quad Q \to L\,S \,\}$$

Given a CFG, $\mathcal{G}' : \mathbb{G} = \langle \Sigma, V, P, S \rangle$ in CNF, we can construct a recognizer $R : \mathbb{G} \to \Sigma^n \to \mathbb{B}$ for strings $\sigma : \Sigma^n$ as follows. Let 2^V be our domain, 0 be $\varnothing$, $\oplus$ be $\cup$, and $\otimes$ be defined as:

$$X \otimes Z := \{\, w \mid \langle x, z \rangle \in X \times Z, (w \to xz) \in P \,\} \tag{1}$$

If we define $\sigma_r^{\pm} := \{w \mid (w \to \sigma_r) \in P\}$, then construct a matrix with nonterminals on the superdiagonal representing each token, $M^0_{r+1=c}(\mathcal{G}', e) := \sigma_r^{\pm}$ and solve for the fixpoint $M^* = M + M^2$,

$$M^0 := \begin{pmatrix} \varnothing & \sigma_1^{\rightarrow} & \varnothing & \cdots & \varnothing \\ & \ddots & \ddots & & \\ & & & & \varnothing \\ & & & & \sigma_n^{\uparrow} \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \begin{pmatrix} \varnothing & \sigma_1^{\rightarrow} & \Lambda & \cdots & \varnothing \\ & \ddots & \ddots & & \\ & & & & \Lambda \\ & & & & \sigma_n^{\uparrow} \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \ldots \Rightarrow M^* = \begin{pmatrix} \varnothing & \sigma_1^{\rightarrow} & \Lambda & \cdots & \Lambda_\sigma^* \\ & \ddots & \ddots & & \\ & & & & \Lambda \\ & & & & \sigma_n^{\uparrow} \\ \varnothing & & & & \varnothing \end{pmatrix}$$

we obtain the recognizer, $R(\mathcal{G}', \sigma) := [S \in \Lambda^*_\sigma] \Leftrightarrow [\sigma \in \mathcal{L}(\mathcal{G})]$ [1].

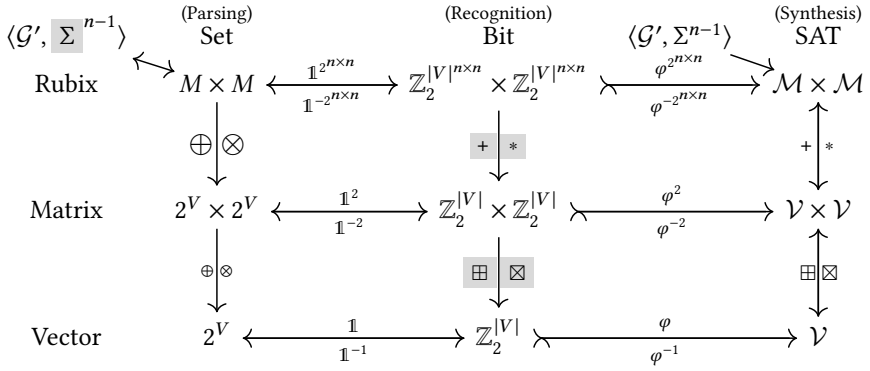
Since $\bigoplus_{c=1}^{n} M_{r,c} \otimes M_{c,r}$ has cardinality bounded by $|V|$, it can be represented as $\mathbb{Z}_2^{|V|}$ using the characteristic function, $\mathbb{1}$. Note that any encoding which respects linearity $\varphi(\Lambda \circledast \Lambda') \equiv \varphi(\Lambda) \circledast \varphi(\Lambda')$ is suitable – this particular representation shares the same algebraic structure, but is more widely studied in error correction, and readily compiled into circuits and BLAS primitives. Furthermore, it enjoys the benefit of complexity-theoretic speedups to matrix multiplication.

Details of the bisimilarity between parsing and matrix multiplication can be found in Valiant [29], who first realized its time complexity was subcubic $\mathcal{O}(n^\omega)$ where ω is the asymptotic lower bound for Boolean matrix multiplication ($\omega < 2.77$), and Lee [20], who shows that speedups to CFL parsing were realizable by Boolean matrix multiplication algorithms. While more efficient

[1] Hereinafter, we shall use Iverson brackets to denote the indicator function of a predicate, i.e., $[P] \Leftrightarrow \mathbb{1}(P)$.

specialized parsers are known to exist for restricted CFGs, this technique is typically linear arithmic under sparsity and believed to be the most efficient general procedure for CFL parsing.

Valiant's decision procedure can be abstracted by lifting into the domain of bitvector variables, i.e., linear equations over finite fields, where each nonterminal inhabitant of the northeasternmost bitvector $\mathcal{T}$ will instead become an algebraic expression whose solutions correspond to valid parse forests for an incomplete string on the superdiagonal. This yields a novel interpretation of Valiant's algorithm as an equational theory over finite fields, allowing us to solve for admissible completions and their parse forests. In particular, $\boxplus$ and $\boxtimes$ are defined so the following diagram commutes,[2]



where $\mathcal{V}$ is a function $\mathbb{Z}_2^{|V|} \to \mathbb{Z}_2$. Note that while always possible to encode $\mathbb{Z}_2^{|V|} \to \mathcal{V}$ using the identity function, an arbitrary $\mathcal{V}$ might have zero, one, or in general, multiple solutions in $\mathbb{Z}_2^{|V|}$. In practice, this means that a language equation can be unsatisfiable or underconstrained, however if a solution exists, it can always be decoded into a valid sentence and parse forest in the language.

So far, we have only considered the syntactic theory of breadth-bounded CFLs with holes, however, our construction can be easily extended to handle the family of CFLs closed under conjunction. The additional expressivity afforded by the language conjunction will be indispensable when considering more practical program repair scenarios that may not be context-free.

4.1 Linear conjunctive reachability

While generally quite expressive, CFLs are themselves not closed under intersection and have other practical limitations, i.e., are unable to express indentation or variable binding. These limitations motivate us towards more expressive yet still efficiently parsable formalisms. In the case of intersection, let us consider the traditional example, $\mathcal{L}_\cap := \mathcal{L}(\mathcal{G}_1) \cap \mathcal{L}(\mathcal{G}_2)$ defined as follows:

$$P_1 := \{\ S \to L\ R, \quad L \to a\ b \mid a\ L\ b, \quad R \to c \mid c\ R\ \}$$
$$P_2 := \{\ S \to L\ R, \quad R \to b\ c \mid b\ R\ c, \quad L \to a \mid a\ L\ \}$$

Note that $\mathcal{L}_\cap$ generates the language $\{\ a^d b^d c^d \mid d > 0\ \}$, which according to the pumping lemma is not context-free. However, we can encode the intersection between two or more languages as a single SAT formula by representing each upper-triangular matrix $\bigcap_{i=1}^{c} \mathcal{L}(\mathcal{G}_i)$ as a polygonal prism with upper-triangular matrices adjoined to each rectangular face. More precisely, we intersect all terminals $\Sigma_\cap := \bigcap_{i=1}^{c} \Sigma_i$, then for each $t_\cap \in \Sigma_\cap$ and CFG, construct an equivalence class $E(t_\cap, \mathcal{G}_i) = \{w_i \mid (w_i \to t_\cap) \in P_i\}$ and bind them together using conjunction:

[2]Hereinafter, we use gray highlighting to denote types and functions defined over strings and binary constants only.

$$\bigwedge_{t\in\Sigma_\cap}\bigwedge_{j=1}^{c-1}\bigwedge_{i=1}^{|\sigma|} E(t_\cap,\mathcal{G}_j) \equiv_{\sigma_i} E(t_\cap,\mathcal{G}_{j+1}) \tag{2}$$

Fig. 2. Orientations of a $\bigcap_{i=1}^{4}\mathcal{L}(\mathcal{G}_i)\cap\Sigma^6$ configuration. As $c\to\infty$, this shape approximates a circular cone whose symmetric axis joins σ_i with orthonormal unit productions $w_i\to t_\cap$, and $[S_i\in\Lambda_\sigma^*]$ represented by the outermost bitvector inhabitants. Equations of this form are equiexpressive with the family of CSLs realizable by finite CFL intersection.

Following Okhotin [22], we extend our grammar DSL with one additional operator for combining CFGs, $\wedge:\mathbb{G}^+\times\mathbb{G}^+\to\mathbb{G}^+$, where $\mathbb{G}^+$ is a conjunctive grammar (CG). In our setting, CGs naturally subsume CFGs, and observe the following denotational semantics:

$$\frac{\Gamma\vdash\mathcal{G}:\mathbb{G}}{\Gamma\vdash\mathcal{G}:\mathbb{G}^+}+\quad\frac{\Gamma\vdash\mathcal{G},\mathcal{G}':\mathbb{G}^+}{\Gamma\vdash\mathcal{G}\wedge\mathcal{G}':\mathbb{G}^+}\wedge\quad\frac{\Gamma\vdash\sigma:\mathcal{L}(\mathcal{G})\quad\Gamma\vdash\sigma:\mathcal{L}(\mathcal{G}')}{\Gamma\vdash\sigma:\mathcal{L}(\mathcal{G}\wedge\mathcal{G}')}\in\quad\frac{\Gamma\vdash\sigma:\mathcal{L}(\mathcal{G}\wedge\mathcal{G}')}{\Gamma\vdash\sigma:\mathcal{L}(\mathcal{G})\cap\mathcal{L}(\mathcal{G}')}\cap$$

Given two CFLs $\mathcal{L}(\mathcal{G}),\mathcal{L}(\mathcal{G}')$, we can compute their intersection $\mathcal{L}(\mathcal{G})\cap\mathcal{L}(\mathcal{G}')\cap\Sigma^d$ by encoding $[\mathbf{M}^*_{\mathcal{G}}\equiv_\sigma\mathbf{M}^*_{\mathcal{G}'}]$. With this feature, we now have the ability to compute the language intersection between two or more CFLs, which will be useful for solving bounded Levenshtein reachability.

4.2 Levenshtein reachability

The Levenshtein ball is finite and therefor context-free, however materializing this set directly is almost always intractable. Instead, we dynamically instantiate a CFG that will recognize and generate the members of $\Delta_d(\underset{\sim}{\sigma})$ for any arbitrary $d:\mathbb{N},\sigma\in\Sigma^*$. This approach follows from a straightforward extension of Levenshtein automata [26].

In the case where σ and σ' are both fixed strings, Levenshtein distance can be interpreted as a shortest path problem over an unweighted graph whose vertices are the strings σ,σ', and all possible intermediate editor states, and edges represent the Levenshtein edits. Thus viewed, the Levenshtein distance between σ and σ' is simply the geodesic distance [23]. When σ' is instead a free variable and the distance is fixed, we can define a finite automaton accepting all and only strings within Levenshtein distance d of σ by unrolling the transition dynamics $\mathcal{L}(\sigma,d)$ up to a fixed horizon d.

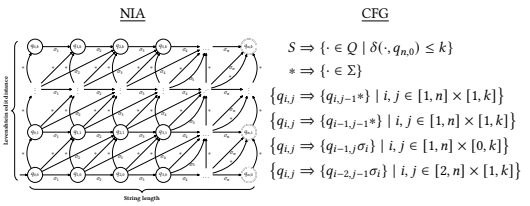


Fig. 3. Levenshtein reachability from Σ^n can be described as an NFA, or a left-linear CFG.

Levenshtein reachability, then, is recognizable by the nondeterministic infinite automaton (NIA) whose topology $\mathcal{L}=$  factorizes into a product of (a) the monotone Chebyshev topology  , equipped with horizontal transitions accepting σ_i and vertical transitions accepting Kleene stars, and (b) the monotone knight's topology  , equipped with transitions accepting σ_{i+2}. The structure of this space is approximated by an acyclic NFA, populated by accept states within radius k of $q_{n,0}$, or equivalently, a left-linear CFG whose productions bisimulate the NFA.

4.3 Bounded Levenshtein Reachability

Let $G(\underset{\sim}{\sigma} : \Sigma^*, d : \mathbb{N}^+) \mapsto \mathbb{G}$ be the specific construction described in Sec. 4.2 which accepts a string, $\underset{\sim}{\sigma}$, and an edit distance, d, and returns a grammar representing the NFA that recognizes the language of all strings within Levenshtein radius d of $\underset{\sim}{\sigma}$. To find the language edit distance and corresponding least-distance edits, we must find the smallest d such that $\mathcal{L}_d^{\cap}$ is nonempty, where $\mathcal{L}_d^{\cap}$ is defined as $\mathcal{L}\big(G(\underset{\sim}{\sigma}, d)\big) \cap \mathcal{L}(\mathcal{G}')$. In other words, we seek $\tilde{\sigma}$ and d^* under which three criteria are all satisfied: (1) $\tilde{\sigma} \in \mathcal{L}(\mathcal{G}')$, and (2) $\Delta(\underset{\sim}{\sigma}, \tilde{\sigma}) \leq d^* \iff \tilde{\sigma} \in \mathcal{L}\big(G(\underset{\sim}{\sigma}, d^*)\big)$, and (3) $\nexists \sigma' \in \mathcal{L}(\mathcal{G}').\big[\Delta(\underset{\sim}{\sigma}, \sigma') < d^*\big]$. To satisfy these criteria, it suffices to check $d \in (1, d^*]$ by encoding the Levenshtein automata and the original grammar as a single SAT formula, call it, $\varphi_d(\cdot)$, and gradually admitting new acceptance states at increasing radii. If $\varphi_d(\cdot)$ returns UNSAT, d is increased until either (1) a satisfying assignment is found or (2) d^* is attained. Following 3.2, this procedure is guaranteed to terminate in at most either (1) the number of steps required to overwrite every symbol in $\underset{\sim}{\sigma}$, or (2) the length of the shortest string in $\mathcal{L}(\mathcal{G}')$, whichever is greater. When $\mathcal{L}(\mathcal{G}')$ is context-free, the language intersection is provably context-free using the Bar-Hillel construction [5], however in general, the resulting intersection is a conjunctive language.

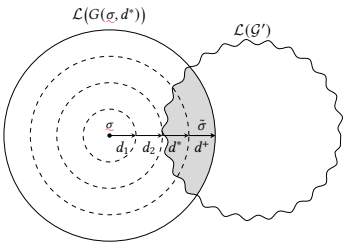


Fig. 4. LED is computed gradually by incrementing d until $\mathcal{L}_d^{\cap} \neq \varnothing$.

4.4 Coarsening and simplification

Depicted right is a SAT tensor representing $\sigma_1\ \sigma_2\ \sigma_3\_ \ldots \_$ where shaded regions demarcate known bitvector literals $\mathcal{L}_{r,c}$ (i.e., representing established nonterminal forests) and unshaded regions correspond to bitvector variables $\mathcal{V}_{r,c}$ (i.e., representing seeded nonterminal forests to be solved) for an incomplete string. Since $\mathcal{L}_{r,c}$ are fixed, we precompute them outside the SAT solver.

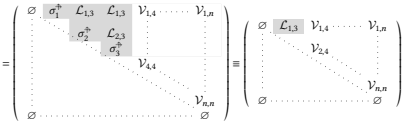


Fig. 5. When is this reduction admissible?

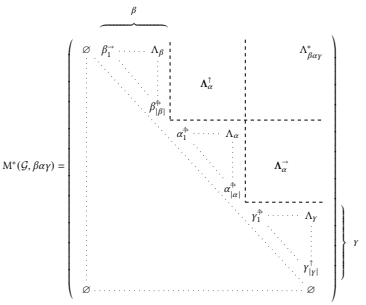


Fig. 6. We can inspect $\mathbf{\Lambda}_\alpha^{\rightarrow}, \mathbf{\Lambda}_\alpha^{\uparrow}$ to find out.

Clearly, solving complexity is heavily dependent on the string length. For the sake of complexity, it would be highly advantageous if well-formed subexpressions could be collapsed into a single nonterminal, or multiple nonterminals (in case of ambiguity), to simplify source code snippets without resorting to ad-hoc program slicing techniques. Naturally, this raises the question of when can partial derivations be simplified, i.e., under what circumstances is the following reduction admissible?

This transformation is admissible when the subexpression is "complete", i.e., its derivation cannot be altered by appending or prepending text. For example, the string `( - b )` is morally *complete* in the sense that inserting adjacent text should not alter the interior derivation, whilst `- b` is not, as introducing adjacent text (e.g., `a - b`) may alter the derivation of its contents depending on the structure of the CFG. This question can be reduced to a quotient: Does there exist another nonterminal, when so adjoined that will "strip away" any tokens, leading to another derivation?

More formally, given an arbitrary (potentially ambiguous) context free grammar $\mathcal{G} : \mathbb{G}$, and string $\alpha : \Sigma^*$, is there a decision procedure that returns whether appending and/or prepending

symbols can alter the parse forest of α? In other words, we want a function $F : (\mathbb{G} \times \Sigma^*) \to \mathbb{B}$ that returns whether α's parse forest according to $\mathcal{G}$ is unique over $\beta\alpha\gamma$, for all $\beta, \gamma : \Sigma^*$.

Specifically, let Λ_α denote the set of all parse trees that are generated by the string α using $\mathcal{G}$, and consider $\mathbf{\Lambda}_\alpha$, the union of all parse trees and their subtrees that (1) can be generated by $\beta\alpha\gamma$ using $\mathcal{G}$ for arbitrary $\beta, \gamma \in \Sigma^*$, and (2) have a leaf in α. We call the parse forest Λ_α *unique* iff $\forall t \in \mathbf{\Lambda}_\alpha \exists t' \in \Lambda_\alpha$, such that t is either a subtree of t', or t' is a proper subtree of t.

4.5 Sampling without replacement

Now that we have a reliable method to synthesize admissible completions for strings containing holes, i.e., fix *localized* errors, $S : \mathcal{G} \times (\Sigma \cup \{\varepsilon, \_\})^n \to \{\Sigma^n\} \subseteq \mathcal{L}(\mathcal{G})$, how can we use S to repair some unparseable string, i.e., $\sigma_1 \dots \sigma_n : \Sigma^n \cap \overline{\mathcal{L}(\mathcal{G})}$ where the holes' locations are unknown? Three questions stand out in particular: how many holes are needed to repair the string, where should we put those holes, and how ought we fill them to obtain a parseable $\tilde{\sigma} \in \mathcal{L}(\mathcal{G})$?

One plausible approach would be to draw samples with a PCFG, minimizing tree-edit distance, however these are computationally expensive metrics and approximations may converge poorly. A more efficient strategy is to sample string perturbations, $\boldsymbol{\sigma} \sim \Sigma^{n \pm q} \cap \Delta_q(\sigma)$, from the Levenshtein q-ball centered on σ, i.e., the space of all admissible edits with Levenshtein distance $\leq q$, loosely analogous to a finite difference approximation over words in a finite language.

More specifically, we employ a pair of [un]tupling functions $\kappa, \rho : \mathbb{N}^k \leftrightarrow \mathbb{N}$ which are (1) bijective (2) maximally compact (3) computationally tractable (i.e., closed form inverses). κ will be used to index $\binom{n}{k}$[2]-combinations via the Maculay representation [19]. ρ will index Σ^k tuples, but is slightly more tricky to define. To maximize compactness, there is an elegant pairing function courtesy of Szudzik [27], which enumerates concentric square shells over the plane $\mathbb{N}^2$ and can be generalized to hypercubic shells in $\mathbb{N}^k$. For our purposes, this generalization will suffice.

Although $\langle \kappa, \rho \rangle$ could be used directly to exhaustively search the Levenshtein ball, they are temporally biased samplers due to lexicographic ordering. Rather, we would prefer a path that uniformly visits every fertile subspace of the Levenshtein ball over time regardless of the grammar or string in question: subsequences of $\langle \kappa, \rho \rangle$ should discover valid repairs with frequency roughly proportional to the filtration rate, i.e., the density of the admissible set relative to the Levenshtein ball. These additional constraints give rise to two more criteria: (1) ergodicity and (2) periodicity.

To achieve ergodicity, we permute the elements of $\binom{n}{k} \times \Sigma^k$ using a finite field with a characteristic polynomial C of degree $m := \lceil \log_p \binom{n}{k} |\Sigma_\varepsilon|^k \rceil$. By choosing C to be some irreducible polynomial, one ensures the path has the mixing properties we desire, e.g., suppose $U : \mathbb{Z}_2^{m \times m}$ is a matrix whose structure is depicted to the right, wherein C represents a primitive polynomial over $\mathbb{Z}_2^m$ with known coefficients $C_{1\dots m}$ and semiring operators $\oplus := + \pmod 2, \otimes := \wedge, \top := 1, \circ := 0$. Since C is primitive, the sequence $\mathbf{S} = (U^{0 \dots 2^m - 1} V)$ must have *full periodicity*, i.e., for all $i, j \in [0, 2^m)$, $\mathbf{S}_i = \mathbf{S}_j \implies i = j$. To uniformly sample $\boldsymbol{\sigma}$ without replacement, we construct a partial surjective function from the finite field onto the Levenshtein ball, $\mathbb{Z}_2^m \rightharpoonup \binom{n}{d} \times \Sigma_\varepsilon^d$, cycle over $\mathbf{S}$, then discard samples which no witness in $\binom{n}{d} \times \Sigma_\varepsilon^d$.

$$U^\intercal V = \begin{pmatrix} C_1 & \cdots & C_m \\ \top & \circ & \circ \\ \circ & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \circ & \circ \; \top & \circ \end{pmatrix}^t \begin{pmatrix} V_1 \\ \vdots \\ \vdots \\ \vdots \\ V_m \end{pmatrix}$$

This procedure requires $\widetilde{\mathcal{O}}(1)$ per sample and roughly $\binom{n}{d} |\Sigma_\varepsilon|^d$ samples to exhaustively search $\binom{n}{d} \times \Sigma_\varepsilon^d$. Its acceptance rate $2^{-m} \binom{n}{d} |\Sigma_\varepsilon|^d$ can be slightly improved with a more suitable base p, however this introduces some additional complexity and so we elected to defer this optimization.

In addition to its statistically desirable properties, our sampler has the practical benefit of being trivially parallelizable using the leapfrog method, i.e., given p independent processors, each one

[2]Following Stirling, we use the notation $\binom{n}{d}$ to denote the set of all d-element subsets of $\{1, \dots, n\}$.

can independently check $[\langle \kappa, \rho \rangle^{-1}(\mathrm{S}_i) \in \mathcal{L}(\mathcal{G})]$ where $i \equiv p_j \pmod{p}$. This procedure linearly scales with the number of processors, exhaustively searching $\Delta_q(\underset{\sim}{\sigma})$ in p^{-1} of the time required by a single processor, or alternately drawing p times as many samples in the same amount of time.

To admit variable-length edits, we first define a ε^+-production and introduce it to the right- and left-hand side of each terminal in a unit production:

$$\frac{\mathcal{G} \vdash \varepsilon \in \Sigma}{\mathcal{G} \vdash (\varepsilon^+ \to \varepsilon \mid \varepsilon\, \varepsilon^+) \in P}\ \varepsilon\text{-DUP} \qquad \frac{\mathcal{G} \vdash (A \to B) \in P}{\mathcal{G} \vdash (A \to B\, \varepsilon^+ \mid \varepsilon^+ B \mid B) \in P}\ \varepsilon^+\text{-INT}$$

Finally, to sample $\boldsymbol{\sigma} \sim \Delta_q(\underset{\sim}{\sigma})$, we enumerate templates $H(\sigma, i) = \sigma_{1\ldots i-1}\ \_\ \_\ \sigma_{i+1\ldots n}$ for each $i \in \cdot \in \binom{n}{d}$ and $d \in 1 \ldots q$, then solve for $\mathcal{M}^*_{\boldsymbol{\sigma}}$. If $[S \in \Lambda^*_{\boldsymbol{\sigma}}]$, then each edit in each $\tilde{\sigma} \in \boldsymbol{\sigma}$ will match one of the following seven patterns:

$$\text{Deletion} = \begin{cases} \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_{1,2} = \varepsilon \end{cases}$$

$$\text{Substitution} = \begin{cases} \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_1 \neq \varepsilon \wedge \gamma_2 = \varepsilon \\ \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_1 = \varepsilon \wedge \gamma_2 \neq \varepsilon \\ \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \{\gamma_1, \gamma_2\} \cap \{\varepsilon, \sigma_i\} = \varnothing \end{cases}$$

$$\text{Insertion} = \begin{cases} \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_1 = \sigma_i \wedge \gamma_2 \notin \{\varepsilon, \sigma_i\} \\ \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_1 \notin \{\varepsilon, \sigma_i\} \wedge \gamma_2 = \sigma_i \\ \ldots \sigma_{i-1}\ \gamma_1\ \gamma_2\ \sigma_{i+1} \ldots & \gamma_{1,2} = \sigma_i \end{cases}$$

This approach is tractable for $n \lesssim 100, q \lesssim 3$, however more complex repairs require a more efficient density estimator. We will now discuss an approach that uses an adaptive sampler (Sec. 4.6) and another based on Levenshtein reachability (Sec. 4.2).

4.6 Probabilistic reachability

Since there are $\Sigma_{d=1}^{q} \binom{n}{d}$ total sketch templates, each with $(|\Sigma| + 1)^d$ individual edits to check, if n and q are large, this space can be intractable to exhaustively search and a uniform prior may be highly sample-inefficient. Furthermore, naïvely sampling $\boldsymbol{\sigma}_i \sim \Sigma^{n \pm q} \cap \Delta_q(\underset{\sim}{\sigma})$ is likely to produce a large number of unnatural edits. To provide rapid and relevant suggestions, we prioritize candidate repairs according to the following seven-step procedure:

(1) Retrieve the most recent grammar, $\mathcal{G}$, and string, $\underset{\sim}{\sigma}$, from the editor.
(2) Sample completions for each template from $\boldsymbol{\sigma}_i \sim \binom{n}{d} \times \Sigma^{n \pm q} \cap \Delta_q(\underset{\sim}{\sigma})$ WoR using Eq. **??**.
(3) Filter completions by admissibility with respect to the grammar, $\boldsymbol{\sigma}_i \cap \mathcal{L}_{\mathcal{G}}$.
(4) Rerank admissible repairs by the edit cost model, $C(\underset{\sim}{\sigma}, \tilde{\sigma})$.
(5) Display the top-k repairs by edit cost found within p-seconds to the user.

Suppose we are given an invalid string, $\underset{\sim}{\sigma} : \Sigma^{90}$ and $\mathcal{F}_\theta$, a distribution over possible edits locations provided by a probabilistic or neural language model, which we can use to localize admissible repairs. For example, by marginalizing onto $\underset{\sim}{\sigma}$, the distribution $\mathcal{F}_\theta$ could take the following form:

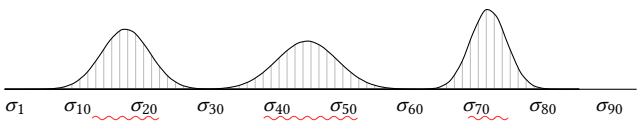


Fig. 7. The distribution $\int \mathcal{F}_\theta(\cdot \mid i_{1\ldots d}) d\theta$, projected onto the invalid string, suggests edit locations most likely to yield admissible repairs, from which we draw subsets of size d.

More specifically in our setting, we want to sample from a discrete product space that factorizes into (1) the specific edit locations (e.g., informed by caret position, historical edit locations, or a static analyzer), (2) probable completions (e.g., from a Markov chain or neural language model) and (3) an accompanying *cost model*, $C : (\Sigma^* \times \Sigma^*) \to \mathbb{R}$, which may be any number of suitable distance metrics, such as language edit distance, finger travel distance on a physical keyboard, weighted Levenshtein distance, or stochastic contextual edit distance [12] in the case of probabilistic edits. Our goal then, is to discover repairs which minimize $C(\sigma, \tilde{\sigma})$, subject to the given grammar and latency constraints.

4.7 Patch minimization

Suppose we have a string, `a ( b`, and discover the patch, $\tilde{\sigma}$ = `( a + b )`. Although $\tilde{\sigma}$ is syntactically admissible, it is not minimal. To minimize a patch, we consider the set of all of its constituent subpatches, namely, `( a + b`, `( a ( b )`, `a + b )`, `( a ( b`, `a + b`, and `a ( b )`, then retain only the smallest syntactically valid instance(s) by Levenshtein distance. This forms a so-called *patch powerset*, which can be lazily enumerated from the top-down using the Maculay representation, after which we take all valid elements from the lowest level containing at least one admissible element, i.e., `a + b` and `a ( b )`. When patches are very large, minimization can be used in tandem with the delta debugging technique [33] to first simplify contiguous edits, then apply the patch powerset construction. Minimization is often useful for estimating the language edit distance: given a single valid repair of arbitrary size, minimization lets us quickly approximate an upper-bound on $\Delta(\sigma, \ell)$ – much tighter than indicated by Lemma 3.2.

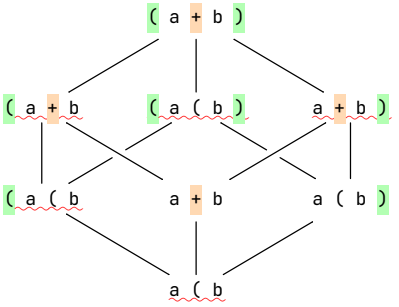


Fig. 8. Patch powerset of $\tilde{\sigma}$ = `( a + b )`.

5 PARSING

Although parsing is not the primary objective of this work, it is an integral component of the repair process, and one of the advantages of using a matrix representation is its excellent error recovery properties. Unlike traditional parsers which fail on a error, matrix-based parsers support parsing invalid strings. In this section, we will describe how to parse strings using matrix powering, and how those results inform the repair process.

5.1 Tree Denormalization

Our parser emits a binary forest consisting of parse trees for the candidate string which are constructed bottom-up using a variant of $\otimes$ called $\hat{\otimes}$, which simply records backpointers:

$$X \,\hat{\otimes}\, Z := \left\{ w \swarrow\!\!\searrow^{x}_{z} \mid \langle x, z \rangle \in X \times Z, (w \to xz) \in P \right\} \tag{3}$$

Due to Chomsky normalization however, the resulting forests are full of trees that are thin and crooked. To restore the natural shape of the tree, we first construct the parse forests bottom-up, then prune away synthetic nonterminals top-down by recursively grafting denormalized grandchildren onto the root. This transformation is purely cosmetic and only used when rendering the parse trees.

Algorithm 1 Tree denormalization

```
procedure CUT(t: Tree)
    stems ← {CUT(c) | c ∈ t.children}
    if t.root ∈ (V_G' \ V_G) then
        return stems
    else
        return { Tree(t.root, stems) }
    end if
end procedure
```

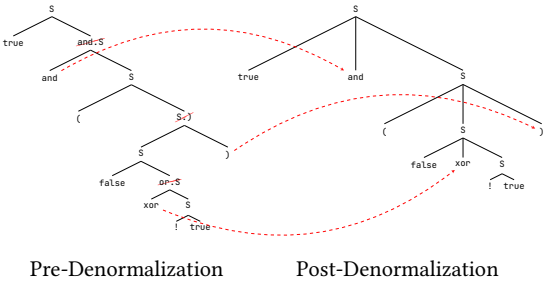


Fig. 9. Since $\mathcal{G}'$ contains synthetic nodes, to recover a parse tree congruent with the original grammar $\mathcal{G}$, we prune all synthetic nodes and graft their stems onto the grandparent via a simple recursive procedure (Alg. 1).

5.2 Relation between parsing and repair

Parsing and repair are intimately related: parsing tells us roughly where the repairs should occur and repair is aided by the results of parsing: parseable subtrees can be used to guide the repair process. Although reparsing may be avoided after repair, it is not straightforward how to decode the solution to the matrix equivalence relation as a parse forest, so we resort to decoding the generating string, then reparsing to obtain the CSTs.

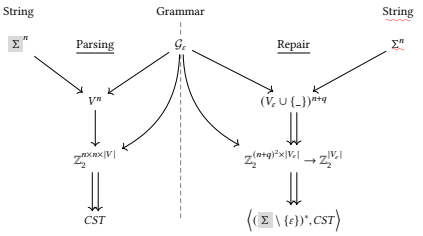


Fig. 10. Parsing and repair share the same grammar representation, i.e., LCGs.

The process of parsing the string $S + S = S$ by matrix powering is depicted below. This occurs by repeatedly joining rooted subtrees using the $\hat{\otimes}$ operator. In this particular case, the parse forest is unique, although in general, M_{ij} may contain multiple subtrees should ambiguity arise:

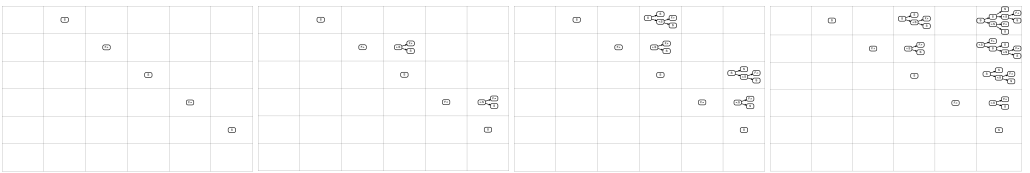


Fig. 11. The parse forest for the string $S + S = S$ is constructed incrementally by computing $M^2 + M = M$.

Our algorithm produces set of concrete syntax trees (CSTs) for a given valid string. Otherwise, if the string is invalid, the algorithm generates a set of admissible corrections, alongside their CSTs.

5.3 Incrementalization

When a parsed string is altered, we can reuse prior work by only recomputing affected submatrices, yielding a reparser whose complexity is location-dependent, i.e., at worst quadratic in terms of $|\Sigma^*|$ assuming $\mathcal{O}(1)$ cost for each CNF-nonterminal subset join, $V_1' \otimes V_2'$. Letting shaded nodes represent observed or bound variables and unshaded nodes as free variables, we depict the worst-case post-editing state of the parse trellis in Fig. 12.

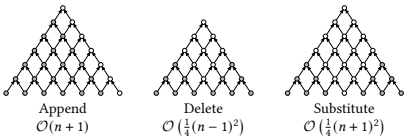


Fig. 12. Incremental reparsing only requires computing submatrices affected by an edit.

The problem of incremental parsing is closely related to *dynamic matrix inversion* in the linear algebra setting, and *incremental transitive closure* with vertex updates in the graph setting. By carefully encoding the matrix relation from Sec. 4 and employing an incremental SAT solver, we can gradually update SAT constraints as new keystrokes are received to eliminate redundancy.

We can use an incremental SAT solver to encode the constraints. To do so, we encode the matrix equivalence relation and use the incremental SAT solver to update the constraints as new keystrokes are received. The incremental SAT solver will eliminate redundancy as it is introduced. Once a solution is found, we can use the algorithm from Sec. 5.1 to recover the parse tree, then introduce a new constraint to block the solution and continue searching for other solutions.

5.4 Error Recovery

Not only is Tidyparse capable of suggesting repairs to invalid strings, it can also return partial trees for those same strings, which is often helpful for debugging purposes. Unlike LL- and LR-style parsers which require special rules for error recovery, Tidyparse can simply analyze the structure of M^* to recover parse branches. If $S \notin \Lambda^*_\sigma$, the upper triangular entries of M^* will take the form of a jagged-shaped ridge whose peaks signify the roots of maximally-parsable substrings $\hat{\sigma}_{i,j}$.

These branches are located on peaks of the upper triangular (UT) matrix ridge. As depicted in Fig. 14, we traverse the peaks by decreasing elevation to collect partial AST branches and display the highest nonoverlapping branches, in this case T_C and T_A to the user, to help them diagnose the parsing error and manually repair it.

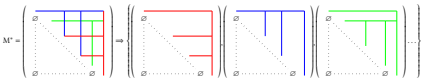


Fig. 13. The matrix $\mathcal{M}^*$ contains all admissible binary trees of a fixed breadth.

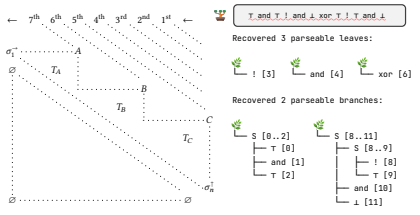


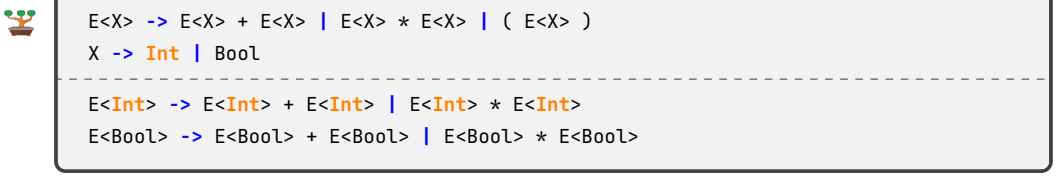
Fig. 14. We can recover the partial subtrees for invalid strings by inspecting the $\mathcal{M}^*$.

5.5 Nonterminal Stubs

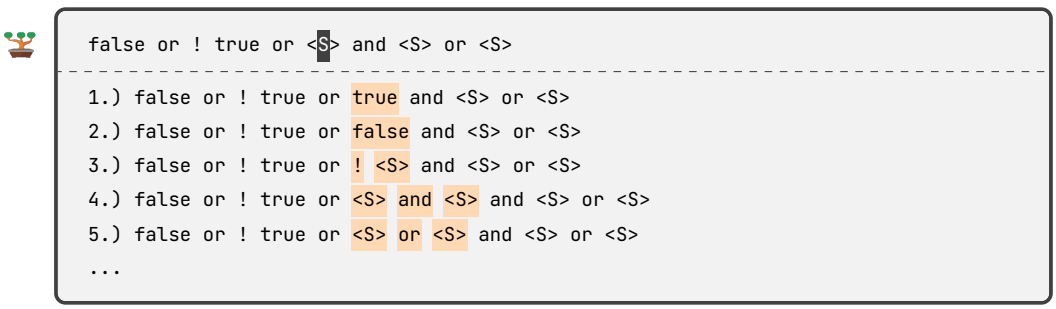
Tidyparse augments CFGs with two additional rules, which are desugared into a vanilla CFG before parsing. The first rule, α-SUB, allows the user to define a nonterminal parameterized by α, a non-recursive nonterminal in the same the CFG representing some finite type and its inhabitants. α-SUB replaces all productions containing $\langle\alpha\rangle$ with the terminals in their transitive closure, $\alpha \rightarrow^* \beta$. The second rule, α-INT, introduces homonymous terminals for each user-defined nonterminal.

$$\frac{\mathcal{G} \vdash (w\langle\alpha\rangle \rightarrow xz) \in P \qquad \alpha^* : \{\beta \mid (\alpha \rightarrow^* \beta) \in P\}}{\mathcal{G} \vdash \forall\beta \in \alpha^*.(w\langle\alpha\rangle \rightarrow xz)[\beta/\alpha] \in P'}\ \alpha\text{-SUB} \qquad \frac{\mathcal{G} \vdash v \in V}{\mathcal{G} \vdash (v \rightarrow \langle v\rangle) \in P}\ \langle\cdot\rangle\text{-INT}$$

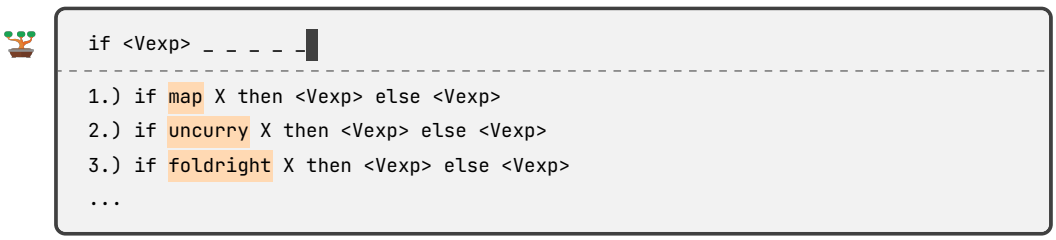
Some dependently typed programming languages can do parsing in the type checker – our parser can also perform some form of type checking. Tidyparse automatically expands typed expressions into ordinary nonterminals using the α-SUB rule, for example when parsing an expression of the form $x + y$, the grammar will recognize `true + false` and `1 + 2`, but not `1 + true`.



When completing a bounded-width string, one finds it is often convenient to admit nonterminal stubs, representing unexpanded subexpressions. To enable this functionality, we introduce a synthetic production for each $v \in V$ using the $\langle\cdot\rangle$-INT rule. Users can interactively build up a complex expression by placing the caret over a stub, then pressing ctrl + Space:

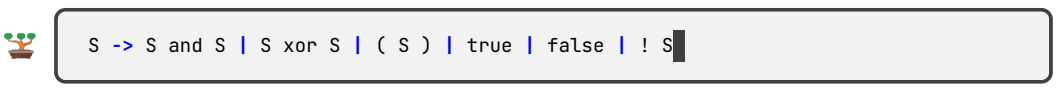


This functionality can also be useful inside a completion, which might be expanded as follows:



6 PRACTICAL EXAMPLE

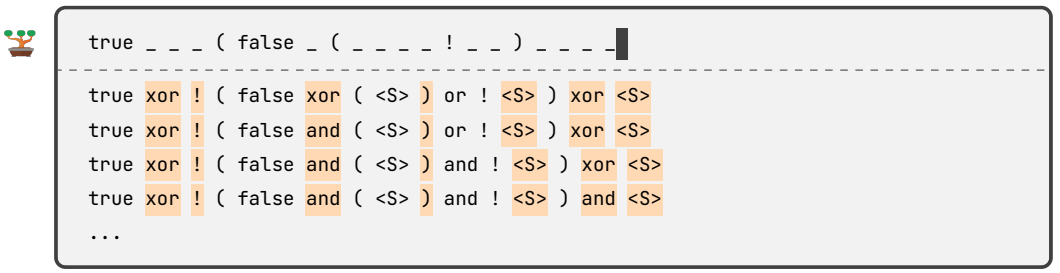
Suppose we are given the following context-free grammar:



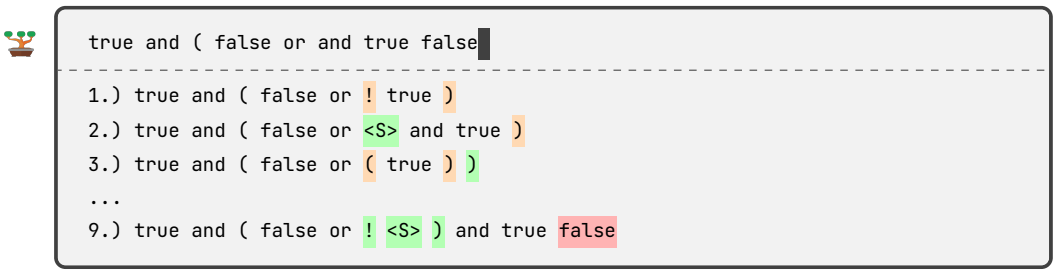
For reasons that will become clear in Sec. 4, this will automatically be rewritten into the grammar:

```
F.! → !      S.) → S F.)   and.S → F.and S   S → F.! S      S → false     S → S ε+
F.( → (    F.xor → xor     xor.S → F.xor S   S → S and.S    S → true      ε+ → ε
F.) → )    F.and → and         S → S xor.S   S → F.( S.)    S → <S>       ε+ → ε+ ε+
```

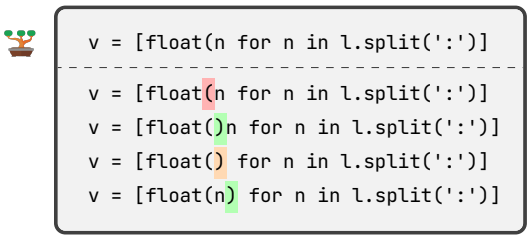
Given a string containing holes, our tool will return several completions in a few milliseconds:



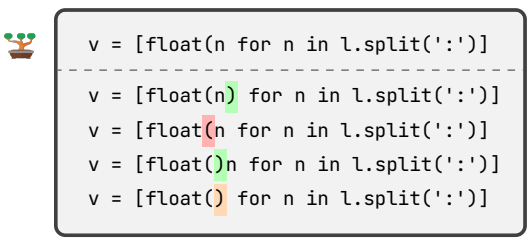
Similarly, if provided with a string containing various errors, it will return several suggestions how to fix it, where green is insertion, orange is substitution and red is deletion.

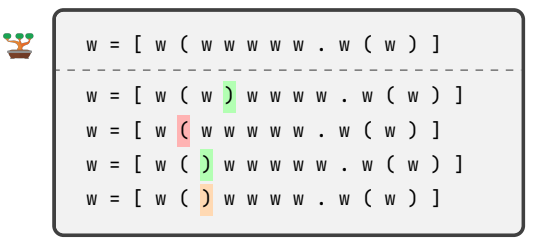


Given an invalid string, the tool will first map the string to a coarsened version and generate edits:

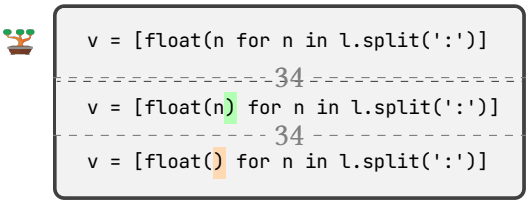


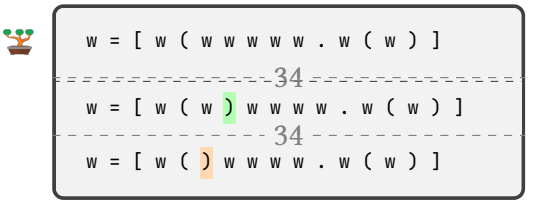
This coarsening is done to reduce the number of possible corrections, and is admissible because CFLs are closed under homomorphisms. (If we wanted to provide a lexical expansion, this would be possible.) These candidates then reranked using a probability metric.



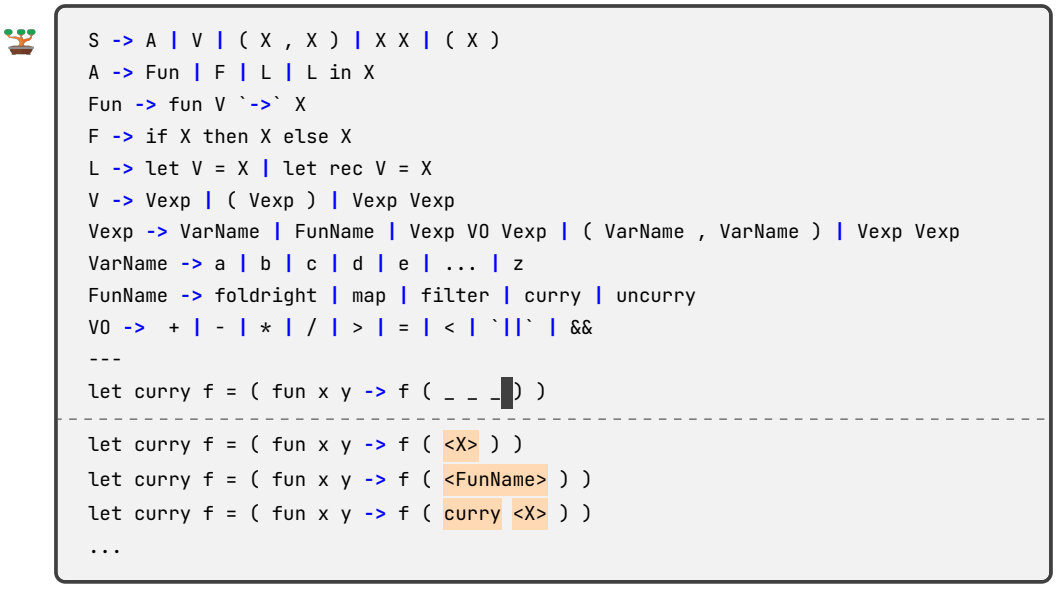


Finally, we eliminate everything which is not parsed by the official Python parser.

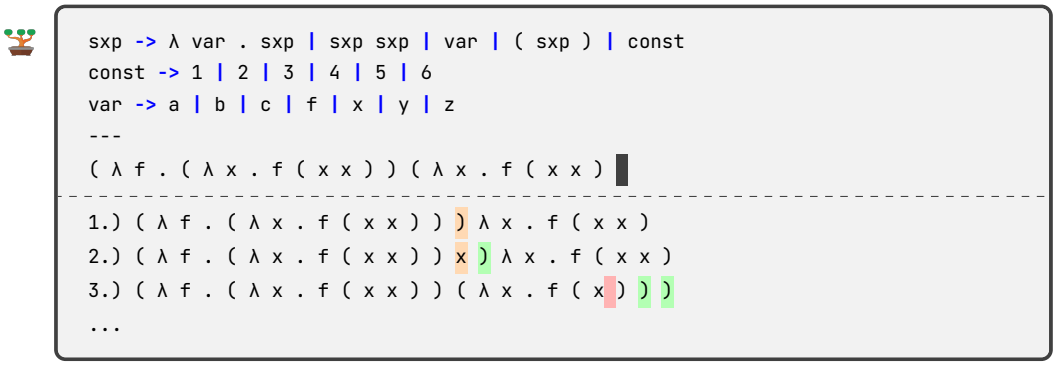




Tidyparse requires a grammar, which can either be provided by the user or ingested from a BNF specification. The following grammar represents a slightly more realistic programming language:



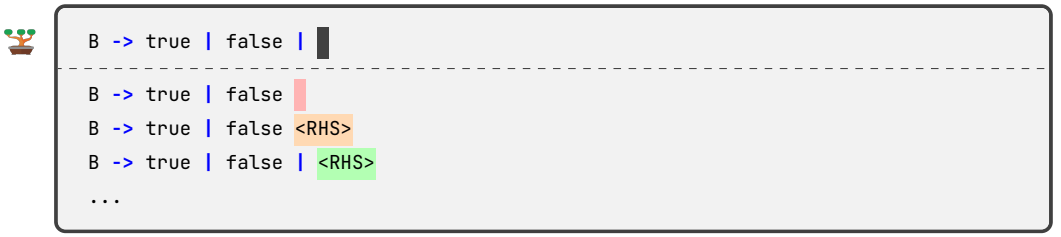
We can also handle error correction and completion in the untyped λ-calculus, as shown below:



Name resolution and scope checking is also possible but requires a more sophisticated grammar.

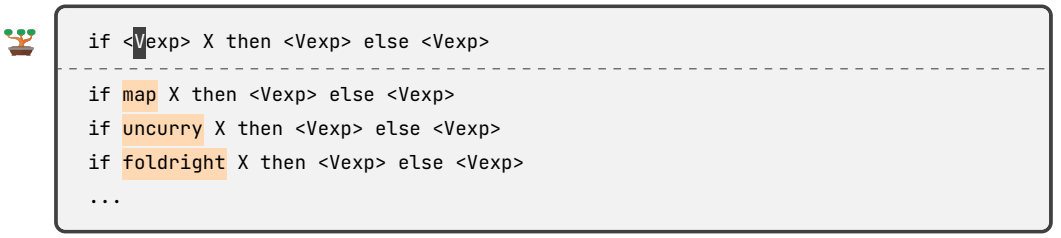
6.1 Grammar Assistance

Tidyparse uses a CFG to parse the CFG, so it can provide assistance while the user is designing the CFG. For example, if the CFG does not parse, it will suggest possible fixes. In the future, we intend to use this functionality to perform example-based codesign and grammar induction.



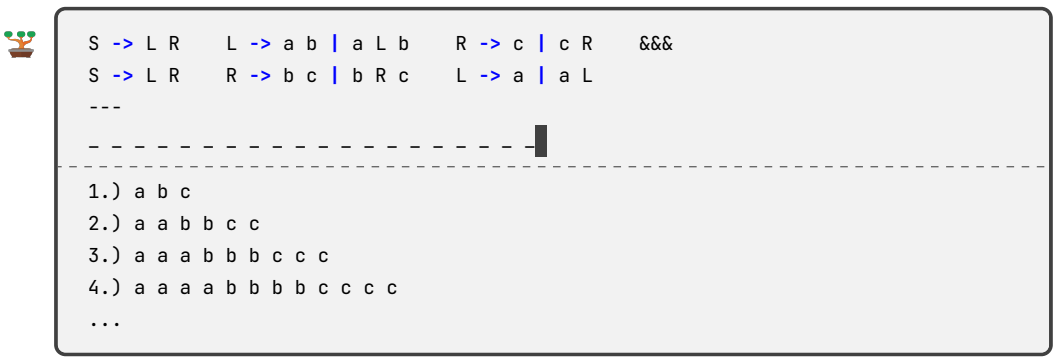
6.2 Interactive Nonterminal Expansion

Users can interactively build up a complex expression by placing the caret over a stub they wish to expand, then pressing ctrl + Space:



6.3 Conjunctive Grammars

Many natural and programming languages exhibit context-sensitivity, such as Python indentation. Unlike traditional parser-generators, Tidyparse can encode CFL intersection, allowing it to detect and correct errors in a more expressive family of languages than would ordinarily be possible using CFGs alone. For example, consider the grammar from Sec. 4.1:



Tidyparse uses the notation `G1 &&& G2` and `G1 ||| G2` to signify $\mathcal{L}_{\mathcal{G}_1} \cap \mathcal{L}_{\mathcal{G}_2}$ and $\mathcal{L}_{\mathcal{G}_1} \cup \mathcal{L}_{\mathcal{G}_2}$ respectively, i.e., the intersection or union of two or more grammars' languages. Composition, complementation and other operations on finite languages are also possible, although undocumented at present.

7 RELATED WORK

Three important questions arise when repairing syntax errors: (1) is the program broken in the first place? (2) if so, where are the errors located? (3) how should those locations then be altered? In the case of syntax correction, those questions are addressed by three related research areas, (1) parsing, (2) language equations and (3) repair. We survey each of those areas in turn.

7.1 CFL Parsing

Context-free language (CFL) parsing is the well-studied problem of how to turn a string into a unique tree, with many different algorithms and implementations (e.g., shift-reduce, recursive-descent, LR). Many of those algorithms expect grammars to be expressed in a certain form (e.g., left- or right- recursive) or are optimized for a narrow class of grammars (e.g., regular, linear).

General CFL parsing allows ambiguity (non-unique trees) and be formulated as a dynamic programming problem, as shown by Cocke-Younger-Kasami (CYK) [24], Earley [15] and others. These parsers have roughly cubic time complexity with respect to the length of the input string.

As shown by Valiant [29], Lee [20] and others, general CFL recognition is in some sense equivalent to binary matrix multiplication, another well-studied combinatorial problem with broad applications, known to be at worst subcubic. This realization unlocks the door to a wide range of complexity-theoretic and practical speedups to CFL recognition and fast general parsing algorithms.

7.2 Language Equations

Language equations are a powerful tool for reasoning about formal languages and their inhabitants. First proposed by Ginsburg et al. [16] for the ALGOL language, language equations are essentially systems of inequalities with variables representing *holes*, i.e., unknown values, in the language or grammar. Solutions to these equations can be obtained using various fixpoint techniques, yielding members of the language. This insight reveals the true algebraic nature of CFLs and their cousins.

Being an algebraic formalism, language equations naturally give rise to a kind of calculus, vaguely reminiscent of Leibniz' and Newton's. First studied by Brzozowski [9, 10] and Antimirov [3], one can take the derivative of a language equation, yielding another equation, which can be interpreted as a kind of continuation or language quotient, returning the suffixes that complete a given prefix. This technique leads to an elegant family of algorithms for incremental parsing [1, 21] and automata minimization [8]. In our setting, differentiation corresponds to code completion.

In this paper, we restrict our attention to language equations over context-free and weakly context-sensitive languages, whose variables coincide with edit locations in the source code of a computer program, and solutions correspond to syntax repairs. Although prior work has studied the use of language equations for parsing [21], to our knowledge they have never previously been considered for the purpose of code completion or syntax error correction.

7.3 Syntax Repair

In finite languages, syntax repair corresponds to spelling correction, a more restrictive and largely solved problem. Schulz and Stoyan [26] construct a finite automaton that returns the nearest dictionary entry by Levenshtein edit distance. Though considerably simpler than syntax correction, their work shares similar challenges and offers insights for handling more general repair scenarios.

When a sentence is grammatically invalid, parsing grows more challenging. Like spelling, the problem is to find the minimum number of edits required to transform an arbitrary string into a syntactically valid one, where validity is defined as containment in a (typically) context-free language. Early work, including Aho [2, 17] propose a dynamic programming algorithm to computes the minimum number of edits required to fix invalid string. Prior work on error correcting parsing

only considers the shortest edit(s), and does not study multiple edits or the Levenshtein ball, in which the problem of actually generating the repairs is not well-posed, as there are often many valid strings that can be obtained within a given number of edits. We instead focus on bounded Levenshtein reachability, which is the problem of finding useful repairs within a fixed Levenshtein distance of the broken string, which requires language intersection.

7.4 Program Synthesis

There is related work on string constraint solving in the constraint programing literature, featuring solvers like CFGAnalyzer and HAMPI [18], which consider bounded context free grammars and intersections thereof. Axelson et al. (2008) [4] has some work on incremental SAT encoding but does not exploit the linear-algebraic structure of parsing nor provide real-time guarantees. Finally, Loris D'Antoni did some great work on *symbolic automata* [13], a generalization of finite automata which allow allow infinite alphabets and symbolic expressions over them. In none of the constraint programming literature we surveyed do any of the approaches employ matrix-based parsing, and therefor do not enjoy the optimality guarantees of Valiant's parser. Our solver can handle CFGs and conjunctive grammars with finite alphabets and does not require any special grammar encoding. The matrix encoding makes it particularly amenable to parallelization.

7.5 Error Correction

Our work focuses on errors arising from human factors in computer programming, in particular *syntax error correction*, which is the problem of fixing partially corrupted programs. Modern research on error correction however can be traced back to the early days of coding theory, when researchers designed *error-correcting codes* (ECCs) to denoise transmission errors induced by external interference, e.g., collision with a high-energy proton, manipulation by an adversary or even typographical mistake. In this context, *code* can be any logical representation for communicating information between two parties (such as a human and a computer), and an ECC is a carefully-designed which ensures that even if some portion of the message should become corrupted, one can still recover the original message by solving a linear system of equations. When designing ECCs, one typically assumes a noise model over a certain event space, such as the Hamming [14, 28] or Levenshtein [6, 7] balls, from which we draw inspiration for our work.

8 EXPERIMENTAL SETUP

To evaluate our model, we primarily use two datasets. We first evaluate it on 5,600 pairs of (broken, fixed) Python code snippets from Wong et al.'s StackOverflow dataset [31] shorter than 40 lexical tokens, whose minimized patch sizes shorter than four lexical tokens ($|\Sigma| = 50, |\underset{\sim}{\sigma}| < 40, \Delta(\underset{\sim}{\sigma}, \ell) < 4$). Minimization uses the Delta debugging [33] technique to isolate the smallest lexical patch that repairs a broken Python snippet.

In the first set of experiments, we uniformly sample without replacement from the Levenshtein edit ball using a LFSR over $\mathbb{Z}_2^n$, and measure the Precision@k of our repair procedure against human repairs of varying edit distances and latency cutoffs. This provides a baseline for the relative density of the admissible set, and an upper bound on the latency of attaining a given precision.

In the second set of experiments, we use an adaptive sampler that stochastically resamples using a Chinese Restaurant Process (CRP). With this technique, repairs are scored and placed into a ranked cache, and new edits are sampled with exponentially decaying probability relative to their perplexity and additive noise is introduced. Initially, the sampler draws uniformly but is gradually biased towards the most likely repairs.

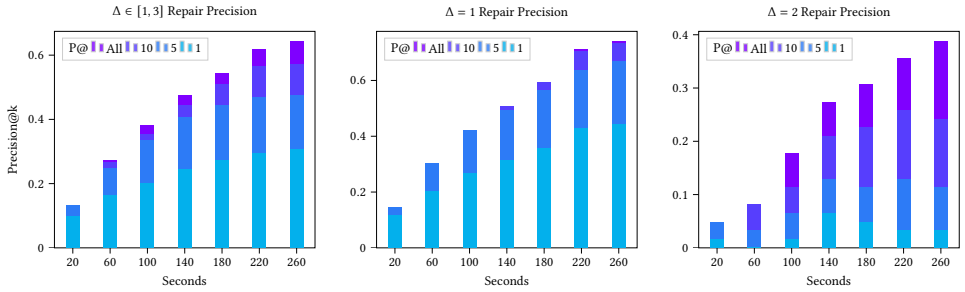
To train the adaptive sampler we use a length-5 variable-order Markov (VOM) chain on 55 million StackOverflow tokens from the BIFI [32] dataset, which takes roughly 10 minutes on 40

Intel Skylake cores running at 2.4 GHz. Sequences are scored using negative log likliehood with Laplace smoothing, or Levenshtein edits when no dataset is available. Evaluation measures the Precision@{1, 5, 10, All} for samples at varying latency cutoffs.

Both sets of experiments use 40 Intel Skylake cores, running at 2.4 GHz, with 16 GB of RAM. We measure the precision on an exact lexical match abstract repair, following the Seq2Parse [25] evaluation, and give a baseline for their approach on the same dataset. Unlike Seq2Parse, our approach does not require a training corpus, and can be used to repair arbitrary context-free and linear conjunctive languages.

8.1 Uniform Sampling Benchmark

Below, we plot the results of a human repair benchmark measuring the Precision@k of our repair procedure against human repairs of varying edit distances and latency cutoffs.



For comparison below are the results from Seq2Parse on the same dataset. Seq2Parse only supports Precision@1 repairs, and so we only report precision@1 from the StackOverflow benchmark for comparison. Unlike our approach which only produces syntactically correct repairs, Seq2Parse also produces syntactically incorrect repairs and so we report the percentage of syntactically correct repairs (Syntactic), as well as the precision of the abstract tokens repairs (AbstractEval), and the exact character match precision (CharMatch).

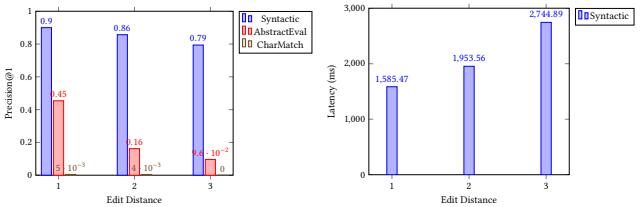


Fig. 15. Seq2Parse precision@1 and latency on the StackOverflow dataset.

8.2 Repair with an adaptive sampler

In the following benchmark, we measure the Precision@k of our repair procedure against human repairs of varying edit distances and latency cutoffs, using an adaptive resampling procedure. This sampler maintains a buffer of successful repairs ranked by perplexity and uses stochastic local search to resample edits within a neighborhood. Initially, edits are sampled uniformly at random. Over time and as the admissible set grows, it prioritizes edits nearby low-perplexity repairs. This technique offers a significant advantage in the low-latency setting.

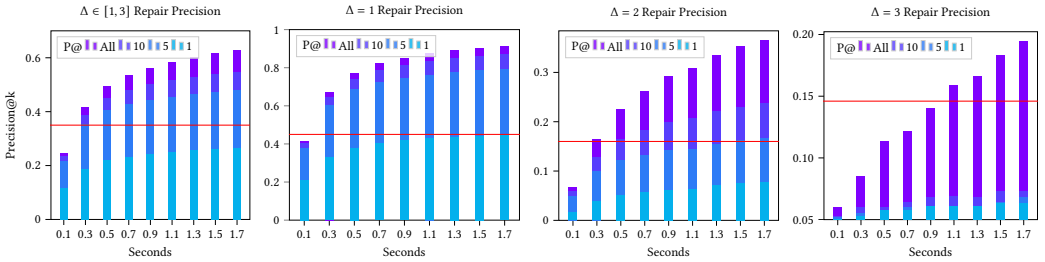


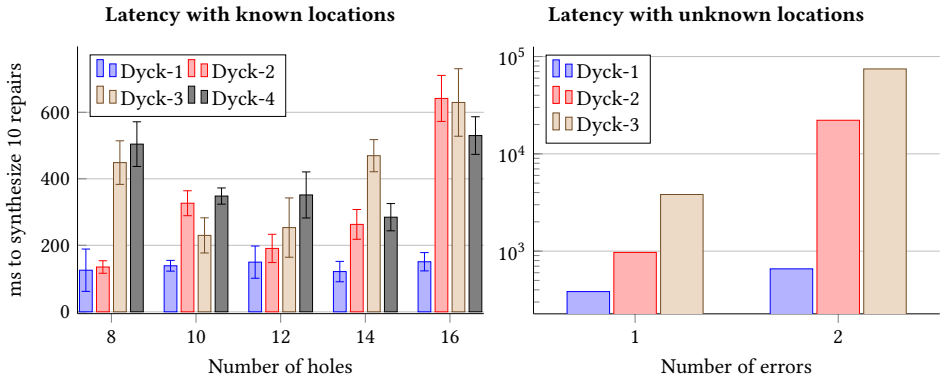
Fig. 16. Adaptive sampling repairs. The red line indicates Seq2Parse precision@1 on the same dataset. Since it only supports generating one repair, we do not report precision@k or the intermediate latency cutoffs.

8.3 Latency Benchmark

In the following benchmarks, we measure the wall clock time required to synthesize solutions to length-50 strings sampled from various Dyck languages, where Dyck-n is the Dyck language containing n different types of balanced parentheses.

```
Dyck-1 -> ( ) | ( Dyck-1 ) | Dyck-1 Dyck-1
Dyck-2 -> Dyck-1 | [ ] | ( Dyck-2 ) | [ Dyck-2 ] | Dyck-2 Dyck-2
Dyck-3 -> Dyck-2 | { } | ( Dyck-3 ) | [ Dyck-3 ] | { Dyck-3 } | Dyck-3 Dyck-3
```

In the first experiment, we sample a random valid string $\sigma \sim \Sigma^{50} \cap \mathcal{L}_{\text{Dyck-n}}$, then replace a fixed number of tokens with holes and measure the average time required to decode ten syntactically-admissible repairs across 100 trial runs.

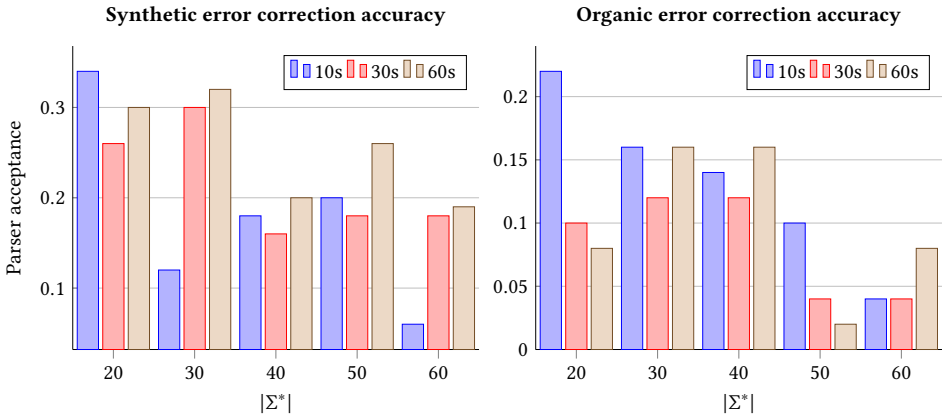


In the second experiment, we sample a random valid string as before, but delete p tokens at random and rather than provide the location(s), ask our model to solve for both the location(s) and repair by sampling uniformly from all n-token HCs, then measure the total time required to decode the first admissible repair. Note the logarithmic scale on the y-axis.

8.4 Accuracy Benchmark

In the following benchmark, we analyze bracketing errors in a dataset of Java and Python code snippets mined from open-source repositories on GitHub. For Java, we sample valid single-line statements with bracket nesting more than two levels deep, synthetically delete one bracket

uniformly at random, repair using Tidyparse[3], then take the top-1 repair after t seconds, and validate using ANTLR's Java 8 parser.



For Python, we sample invalid code fragments uniformly from the imbalanced bracket category of the Break-It-Fix-It (BIFI) dataset [32], a dataset of organic Python errors, repair using Tidyparse, take the top-1 repair after t seconds, and validate repairs using Python's `ast.parse()` method.

9 DISCUSSION

While error correction with a few errors is tolerable, latency can vary depending on many factors including string length and grammar size. If errors are localized to the beginning or end of a string, then latency is typically below 500ms. We observe that errors are typically concentrated nearby historical edit locations, which can be retrieved from the IDE or version control.

Tidyparse in its current form has a number of technical shortcomings: firstly it does not incorporate any neural language modeling technology at present, an omission we hope to address in the near future. Training a language model to predict likely repair locations and rank admissible results could lead to lower overall latency and more natural repairs.

Secondly, our current method generates sketch templates using a naïve enumerative search, feeding them individually to the SAT solver, which has the tendency to duplicate prior work and introduces unnecessary thrashing. Considering recent extensions of Boolean matrix-based parsing to linear context-free rewriting systems (LCFRS) [11], it may be feasible to search through these edits within the SAT solver, leading to yet unrealized and possibly significant speedups.

Lastly and perhaps most significantly, Tidyparse does not incorporate any semantic constraints, so its repairs whilst syntactically admissible, are not guaranteed to be semantically valid. We note however, that it is possible to encode type-based semantic constraints into the solver and intend to explore this direction more fully in future work.

Not only is linear algebra over finite fields an expressive language for inference, but also an efficient framework for inference on languages themselves. We illustrate a few of its applications for parsing incomplete strings and repairing syntax errors in context- free and sensitive languages. In contrast with LL and LR-style parsers, our technique can recover partial forests from invalid strings by examining the structure of M^* and handles arbitrary context-free languages. In future work, we hope to extend our method to more natural grammars like PCFG and LCFRS.

[3]Using the Dyck-n grammar augmented with `D1` → `w` | `D1`. Contiguous non-bracket tokens are substituted with a single placeholder token, `w`, and restored verbatim after bracket repair.

We envision three primary use cases: (1) helping novice programmers become more quickly familiar with a new programming language (2) autocorrecting common typos among proficient but forgetful programmers and (3) as a prototyping tool for PL designers and educators. Featuring a grammar editor and built-in SAT solver, Tidyparse helps developers navigate the language design space, visualize syntax trees, debug parsing errors and quickly generate simple examples and counterexamples for testing.

10 CONCLUSION

The great compromise in program synthesis is one of efficiency versus expressiveness: the more expressive a language, the more concise and varied the programs it can represent, but the harder those programs are to synthesize without resorting to domain-specific heuristics. Likewise, the simpler a language is to synthesize, the weaker its concision and expressive power.

Most existing work on program synthesis focus on general λ-calculi, or very narrow languages such as finite automata or regular expressions. The former are too expressive to be synthesized or verified, whilst the latter are too restrictive to be useful. In our work, we focus on context-free and mildly context-sensitive grammars, which are expressive enough to capture a variety of useful programming language features, but not so expressive as to be unsynthesizable.

The second great compromise in program synthesis is that of reusability versus specialization. In programming, as in human affairs, there is a vast constellation of languages, each requiring specialized generators and interpreters. Are these languages truly irreconcilable? Or, as Noam Chomsky argues, are these merely dialects of a universal language? *Synthesis* then, might be a misnomer, and more aptly called *recognition*, in the analytic tradition.

In our work, we argue these two compromises are not mutually exclusive, but complementary and reciprocal. Programs and the languages they inhabit are indeed synthetic, but can be analyzed and reused in the metalanguage of context-free grammars closed under conjunction. Not only does this admit an efficient synthesis algorithm, but allows users to introduce additional constraints without breaking compositionality, one of the most sacred tenets in programming language design.

Furthermore, we argue it is possible to improve the efficiency of human programmers without sacrificing expressiveness by considering latency to synthesize an acceptable completion. In contrast with program synthesizers that require intermediate programs to be well-formed, our synthesizer is provably sound and complete up to a Levenshtein distance bound, and attempts to minimize total edits, but does not impose any constraints on the code itself being written.

Tidyparse accepts a CFG and a string to parse. If the string is valid, it returns the parse forest, otherwise, it returns a set of repairs, ordered by their Levenshtein edit distance to the invalid string. Our method compiles each CFG and candidate string onto a matrix dynamical system using an extended version of Valiant's construction and solves for its fixedpoints using an incremental SAT solver. This approach to parsing has many advantages, enabling us to repair syntax errors, correct typos and generate parse trees for incomplete strings. By allowing the string to contain holes, repairs can contain either concrete tokens or nonterminals, which can be manually expanded by the user or a neural-guided search procedure. From a theoretical standpoint, this technique is particularly amenable to neural program synthesis and repair, naturally integrating with the masked-language-modeling task (MLM) used by transformer-based neural language models.

From a practical standpoint, we have implemented our approach as an IDE plugin and demonstrated its viability as a tool for live programming. Tidyparse is capable of generating repairs for invalid code in a range of practical languages. We plan to continue expanding its grammar and autocorrection functionality to cover a broader range of languages and hope to conduct a more thorough user study to validate its effectiveness in the near future.

REFERENCES

[1] Michael D Adams, Celeste Hollenbeck, and Matthew Might. 2016. On the complexity and performance of parsing with derivatives. In Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation. 224–236.

[2] Alfred V Aho and Thomas G Peterson. 1972. A minimum distance error-correcting parser for context-free languages. SIAM J. Comput. 1, 4 (1972), 305–312.

[3] Valentin Antimirov. 1996. Partial derivatives of regular expressions and finite automaton constructions. Theoretical Computer Science 155, 2 (1996), 291–319.

[4] Roland Axelsson, Keijo Heljanko, and Martin Lange. 2008. Analyzing context-free grammars using an incremental SAT solver. In Automata, Languages and Programming: 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part II 35. Springer, 410–422.

[5] Yehoshua Bar-Hillel, Micha Perles, and Eli Shamir. 1961. On formal properties of simple phrase structure grammars. Sprachtypologie und Universalienforschung 14 (1961), 143–172.

[6] Daniella Bar-Lev, Tuvi Etzion, and Eitan Yaakobi. 2021. On Levenshtein Balls with Radius One. In 2021 IEEE International Symposium on Information Theory (ISIT). 1979–1984. https://doi.org/10.1109/ISIT45174.2021.9517922

[7] Leonor Becerra-Bonache, Colin de La Higuera, Jean-Christophe Janodet, and Frédéric Tantini. 2008. Learning Balls of Strings from Edit Corrections. Journal of Machine Learning Research 9, 8 (2008).

[8] Janusz A Brzozowski. 1962. Canonical regular expressions and minimal state graphs for definite events. In Proc. Symposium of Mathematical Theory of Automata. 529–561.

[9] Janusz A Brzozowski. 1964. Derivatives of regular expressions. Journal of the ACM (JACM) 11, 4 (1964), 481–494.

[10] Janusz A. Brzozowski and Ernst Leiss. 1980. On equations for regular languages, finite automata, and sequential networks. Theoretical Computer Science 10, 1 (1980), 19–35.

[11] Shay B Cohen and Daniel Gildea. 2016. Parsing linear context-free rewriting systems with fast matrix multiplication. Computational Linguistics 42, 3 (2016), 421–455.

[12] Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2014. Stochastic Contextual Edit Distance and Probabilistic FSTs. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 2 (Short Papers). Association for Computational Linguistics, Baltimore, Maryland, 625–630.

[13] Loris D'Antoni and Margus Veanes. 2014. Minimization of symbolic automata. In Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. 541–553.

[14] Dingding Dong, Nitya Mani, and Yufei Zhao. 2023. On the number of error correcting codes. Combinatorics, Probability and Computing (2023), 1–14. https://doi.org/10.1017/S0963548323000111

[15] Jay Earley. 1970. An efficient context-free parsing algorithm. Commun. ACM 13, 2 (1970), 94–102.

[16] Seymour Ginsburg and H Gordon Rice. 1962. Two families of languages related to ALGOL. Journal of the ACM (JACM) 9, 3 (1962), 350–371.

[17] E. T. Irons. 1963. An Error-Correcting Parse Algorithm. Commun. ACM 6, 11 (nov 1963), 669–673. https://doi.org/10.1145/368310.368385

[18] Adam Kiezun, Vijay Ganesh, Philip J Guo, Pieter Hooimeijer, and Michael D Ernst. 2009. HAMPI: a solver for string constraints. In Proceedings of the eighteenth international symposium on Software testing and analysis. 105–116.

[19] D. E. Knuth. 2005. Generating All Combinations and Partitions. Addison-Wesley, 5–6.

[20] Lillian Lee. 2002. Fast context-free grammar parsing requires fast boolean matrix multiplication. Journal of the ACM (JACM) 49, 1 (2002), 1–15. https://arxiv.org/pdf/cs/0112018.pdf

[21] Matthew Might, David Darais, and Daniel Spiewak. 2011. Parsing with derivatives: a functional pearl. ACM sigplan notices 46, 9 (2011), 189–195.

[22] Alexander Okhotin. 2001. Conjunctive grammars. Journal of Automata, Languages and Combinatorics 6, 4 (2001), 519–535.

[23] Perrin E. Ruth and Manuel E. Lladser. 2023. Levenshtein graphs: Resolvability, automorphisms determining sets. Discrete Mathematics 346, 5 (2023), 113310. https://doi.org/10.1016/j.disc.2022.113310

[24] Itiroo Sakai. 1961. Syntax in universal translation. In Proceedings of the International Conference on Machine Translation and Applied Language Analysis.

[25] Georgios Sakkas, Madeline Endres, Philip J Guo, Westley Weimer, and Ranjit Jhala. 2022. Seq2Parse: neurosymbolic parse error repair. Proceedings of the ACM on Programming Languages 6, OOPSLA2 (2022), 1180–1206.

[26] Klaus U Schulz and Stoyan Mihov. 2002. Fast string correction with Levenshtein automata. International Journal on Document Analysis and Recognition 5 (2002), 67–85.

[27] Matthew Szudzik. 2006. An elegant pairing function. In Special NKS 2006 Wolfram Science Conference. 1–12.

[28] Michalis K Titsias and Christopher Yau. 2017. The Hamming ball sampler. J. Amer. Statist. Assoc. 112, 520 (2017), 1598–1611.

[29] Leslie G Valiant. 1975. General context-free recognition in less than cubic time. Journal of computer and system sciences 10, 2 (1975), 308–315. http://people.csail.mit.edu/virgi/6.s078/papers/valiant.pdf
[30] Leslie G Valiant. 1979. Completeness classes in algebra. In Proceedings of the eleventh annual ACM symposium on Theory of computing. 249–261. https://dl.acm.org/doi/pdf/10.1145/800135.804419
[31] Alexander William Wong, Amir Salimi, Shaiful Chowdhury, and Abram Hindle. 2019. Syntax and Stack Overflow: A methodology for extracting a corpus of syntax errors and fixes. In 2019 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, 318–322.
[32] Michihiro Yasunaga and Percy Liang. 2021. Break-it-fix-it: Unsupervised learning for program repair. In International Conference on Machine Learning. PMLR, 11941–11952.
[33] Andreas Zeller. 2002. Isolating cause-effect chains from computer programs. ACM SIGSOFT Software Engineering Notes 27, 6 (2002), 1–10.

A EXAMPLE REPAIRS

We give some example human repairs that were correctly predicted from the StackOverflow dataset.

Original method	Fixed Repair

`form sympy import *`
`x = Symbol('x', real=True)`
`x, re(x), im(x)`
	`from sympy import *`
	`x = Symbol('x', real=True)`
	`x, re(x), im(x)`

`result = yeald From(item.create())` `raise Return(result)`	`result = yield From(item.create())` `raise Return(result)`
`return 1/sum_p if sum_p \` `return 0 else`	`return 1/sum_p if sum_p \` `else 0`

`from itertools improt permutations, product`
`a='ABC'`
`b=['*', '%3A']`
`l=[a]*2+[b]`
`def g(list):`
`  for p in permutations(list):`
`  yield product(*p)`
`result=g(l)`
	`from itertools import permutations, product`
	`a='ABC'`
	`b=['*', '%3A']`
	`l=[a]*2+[b]`
	`def g(list):`
	`  for p in permutations(list):`
	`  yield product(*p)`
	`result=g(l)`

`sum(len(v) for v items.values()))`	`sum(len(v) for v in items.values())`
`df.apply(lambda row: list(set(row['ids'))))`	`df.apply(lambda row: list(set(row['ids'])))`
`import numpy ad np` `A_concate = np.array([a_0, a_1, a_2,..., a_n])`	`import numpy as np` `A_concate = np.array([a_0, a_1, a_2,..., a_n])`