Syntax Repair as Language Intersection

ANONYMOUS AUTHOR(S)

We introduce a new technique for correcting syntax errors in arbitrary context-free languages. Our work comes from the observation that syntax errors with a small repair typically have very few unique small repairs, which can usually be enumerated up to a small edit distance then ranked within a short amount of time. Furthermore, we place a heavy emphasis on precision: the enumerated set must contain every possible repair within a few edits and no invalid repairs. To do so, we reduce CFL recognition onto Boolean tensor completion, then model error correction as a language intersection problem between a Levenshtein automaton and a context-free grammar. To decode the solutions, we then sample trees without replacement from the intersection grammar, which yields valid repairs within a certain Levenshtein distance. Finally, we rank all repairs discovered within 60 seconds by a Markov chain.

1 INTRODUCTION

Syntax repair is the problem of modifying an invalid sentence so it conforms to some grammar. Prior work has been devoted to fixing syntax errors using handcrafted heuristics. This work features a variety of approaches including rule-based systems and statistical language models. However, these techniques are often brittle, and are susceptible to misgeneralization. In a prior paper published in SPLASH, the authors sample production rules from an error correcting grammar. While theoretically sound, this technique is incomplete, i.e., not guaranteed to sample all edits within a certain Levenshtein distance, and no more. In this paper, we demonsrate it is possible to attain a significant advantage by synthesizing and scoring all repairs within a certain Levenshtein distance. Not only does this technique guarantee perfect generalization, but also helps with precision.

We take a first-principles approach making no assumptions about the sentence or grammar and focuses on correctness and end-to-end latency. Our technique is simple:

(1) We first reduce the problem of CFL recognition to Boolean tensor completion, then use that to compute the Parikh image of the CFL. This follows from a straightforward extension of the Chomsky-Schützenberger enumeration theorem.
(2) We then model syntax correction as a language intersection problem between a Levenshtein automaton and a context-free grammar, which we explicitly materialize using a specialized version of the Bar-Hillel construction to Levenshtein intersections. This greatly reduces the number of generated productions.
(3) To decode the members from the intersection grammar, we sample trees without replacement by constructing a bijection between syntax trees and the integers, then sampling integers uniformly without replacement from a finite range. This yields concrete repairs within a certain Levenshtein distance.
(4) Finally, we rerank all repairs found before a fixed timeout by n-gram perplexity.

Though simple, this technique outperforms SoTA syntax repair techniques. Its efficacy owes to the fact it does not sample edits or nor productions, but unique, fully formed repairs within a certain Levenshtein distance. It is sound and complete up to a Levenshtein bound - i.e., it will find all repairs within an arbitrary Levenshtein distance, and no more. Often the langauge of small repairs is surpisingly small compared with the langauge of all possible edits, enabling us to efficiently synthesize and score every possible solution. This offers a significant advantage over memoryless sampling techniques.

SPLASH'24, October 22-27, 2024, Pasadena, California, United States
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 EXAMPLE

Consider the following Python snippet, which contains a small syntax error:

```
def prepend(i, k, L=[]) n and [prepend(i - 1, k, [b] + L) for b in range(k)]
```

We can fix it by inserting a colon after the function definition, yielding:

```
def prepend(i, k, L=[]): n and [prepend(i - 1, k, [b] + L) for b in range(k)]
```

A careful observer will note that there is only one way to repair this Python snippet by making a single edit. In fact, many programming languages share this curious property: syntax errors with a small repair have few uniquely small repairs. Valid sentences corrupted by a few small errors rarely have many small corrections. We call such sentences *metastable*, since they are relatively stable to small perturbations, as likely to be incurred by a careless typist or novice programmer.

Let us consider a slightly more ambiguous error: `v = df.iloc(5:, 2:)`. Assuming an alphabet of just one hundred lexical tokens, this tiny statement has millions of possible two-token edits, yet only six of those possibilities are accepted by the Python parser:

(1) `v = df.iloc(5:, 2,)` (3) `v = df.iloc(5[:, 2:])` (5) `v = df.iloc[5:, 2:]`

(2) `v = df.iloc(5), 2()` (4) `v = df.iloc(5:, 2:)` (6) `v = df.iloc(5[:, 2])`

With some typing information, we could easily narrow the results, but even in the absence of semantic constraints, one can probably rule out (2, 3, 6) given that `5[` and `2(` are rare bigrams in Python, and knowing `df.iloc` is often followed by `[`, determine (5) is most natural. This is the key insight behind our approach: we can usually locate the intended fix by exhaustively searching small repairs. As the set of small repairs is itself often small, if only we had some procedure to distinguish valid from invalid patches, the resulting solutions could be simply ranked by naturalness.

The trouble is that any such procedure must be highly sample-efficient. We cannot afford to sample the universe of possible d-token edits, then reject invalid ones – assuming it takes just 10ms to generate and check each sample, (1-6) could take 24+ hours to find. The hardness of brute-force search grows superpolynomially with edit distance, sentence length and alphabet size. We need a more efficient procedure for sampling all and only small valid repairs.

3 PROBLEM

We can model syntax repair as a language intersection problem between a context-free language (CFL) and a regular language.

Definition 3.1 (Bounded Levenshtein-CFL reachability). Given a CFL ℓ and an invalid string $\underset{\sim}{\sigma} : \ell^{\complement}$, the BCFLR problem is to find every valid string reachable within d edits of $\underset{\sim}{\sigma}$, i.e., letting Δ be the Levenshtein metric and $L(\underset{\sim}{\sigma}, d) \coloneqq \{\sigma \mid \Delta(\underset{\sim}{\sigma}, \sigma) \leq d\}$, we seek to find $A = L(\underset{\sim}{\sigma}, d) \cap \ell$.

As the admissible set A is typically under-determined, we want a procedure which prioritizes natural and valid repairs over unnatural but valid repairs:

Definition 3.2 (Ranked repair). Given a finite language $A = L(\underset{\sim}{\sigma}, d) \cap \ell$ and a probabilistic language model $\mathrm{P}_\theta : \Sigma^* \to [0, 1] \subset \mathbb{R}$, the ranked repair problem is to find the top-k maximum likelihood repairs under the language model. That is,

$$R(A, P_\theta) \coloneqq \underset{\{\boldsymbol{\sigma} \mid \boldsymbol{\sigma} \subseteq A, |\boldsymbol{\sigma}| \leq k\}}{\operatorname{argmax}} \sum_{\sigma \in \boldsymbol{\sigma}} \mathrm{P}_\theta(\sigma \mid \underset{\sim}{\sigma}) \tag{1}$$

The problem of ranked repair is typically solved by learning a distribution over strings, however this approach can be sample-inefficient and generalize poorly to new langauges. Representing the problem as a distribution over Σ^* forces the language model to jointly learn syntax and stylometry. As our work demonstrates, this problem can be factorized into a bilevel objective: first retrieval, then ranking. By ensuring retrieval is sufficiently precise and exhaustive, maximizing likliehood over the admissible set can be achieved with a much simpler, syntax-oblivious language model.

Even with an extremely efficient approximate sampler for $\sigma \sim \ell_\cap$, due to the size of ℓ and $L(\underset{\sim}{\sigma}, d)$, it would be intractable to sample either ℓ or $L(\underset{\sim}{\sigma}, d)$, then reject invalid ($\sigma \notin \ell$) or unreachable ($\sigma \notin L(\underset{\sim}{\sigma}, d)$) edits, and completely out of the question to sample $\sigma \sim \Sigma^*$ as do many large language models. Instead, we will explicitly construct a grammar generating $\ell \cap L(\underset{\sim}{\sigma}, d)$, sample from it, then rerank all repairs after a fixed timeout. So long as $|\ell_\cap|$ is sufficiently small and recognizes all and only small repairs, our sampler is sure to retrieve the most natural repair and terminate quickly.

4 METHOD

The syntax of most programming languages is context-free. Our proposed method is simple. We construct a context-free grammar representing the intersection between the langauge syntax and an automaton recognizing the Levenshtein ball of a given radius. Since CFLs are closed under intersection with regular languages, this is admissible. Three outcomes are possible:

(1) $G_\cap$ is empty, in which case there is no repair within the given radius. In this case, we simply increase the radius and try again.
(2) $\mathcal{L}(G_\cap)$ is small, in which case we enumerate all possible repairs. Enumeration is tractable for $\sim 80\%$ of the dataset in ≤ 90s.
(3) $\mathcal{L}(G_\cap)$ is too large to enumerate, so we sample from the intersection grammar $G_\cap$. Sampling is necessary for $\sim 20\%$ of the dataset.

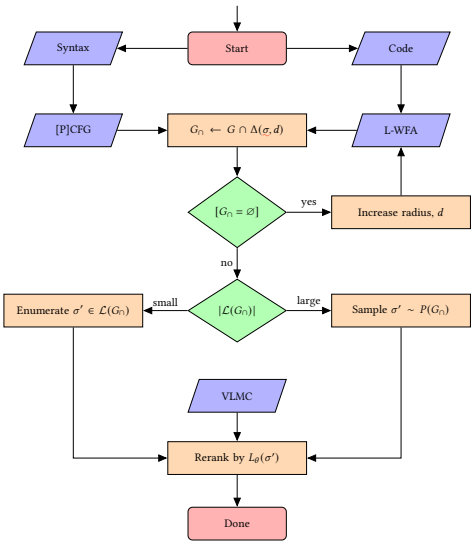


Fig. 1. Flowchart of our proposed method.

As long as we have done our job correctly, the intersection language should contain every repair within a certain Levenshtein distance, and no invalid repairs. This procedure is depicted in Fig. 1. In the following section, we will describe how we construct the intersecection grammar (§ 4.2, 4.3), then, provide an explicit technique for extracting all repairs contained within it (§ 4.5). Finally, we use an n-gram model to rank and return the top-k results by likelihood (§ 4.6).

4.1 Preliminaries

Recall that a CFG, $\mathcal{G} = \langle \Sigma, V, P, S \rangle$, is a quadruple consisting of terminals (Σ), nonterminals (V), productions ($P \colon V \to (V \mid \Sigma)^*$), and a start symbol, ($S$). 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:

$$G \coloneqq \{\, S \to S\,S \mid (\,S\,) \mid (\,) \,\} \Longrightarrow \{\, S \to Q\,R \mid S\,S \mid L\,R, \quad R \to), \quad L \to (, \quad Q \to L\,S \,\}$$

Likewise, a finite state automaton is a quintuple $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$, where Q is a finite set of states, Σ is a finite alphabet, $\delta \subseteq Q \times \Sigma \times Q$ is the transition function, and $I, F \subseteq Q$ are the set of initial and final states, respectively. We will use this notation in the following sections.

4.2 The Nominal Levenshtein Automaton

Levenshtein edits are recognized by an automaton known as the Levenshtein automaton. As the original construction defined by Schultz and Mihov [34] contains cycles and ε-transitions, we propose a variant which is ε-free and acyclic. Furthermore, we adpot a nominal form which supports infinite alphabets and considerably simplifies the langauge intersection. Illustrated in Fig. 2 is an example of a small Levenshtein automaton recognizing $\Delta(\sigma : \Sigma^5, 3)$. Unlabeled arcs accept any terminal from the alphabet, Σ.

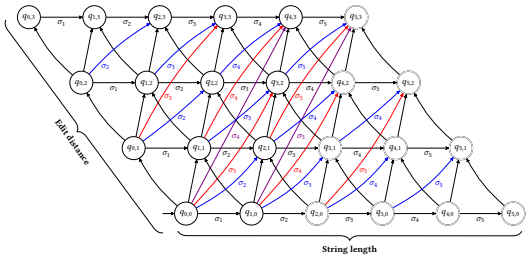


Fig. 2. NFA recognizing Levenshtein $\Delta(\sigma : \Sigma^5, 3)$.

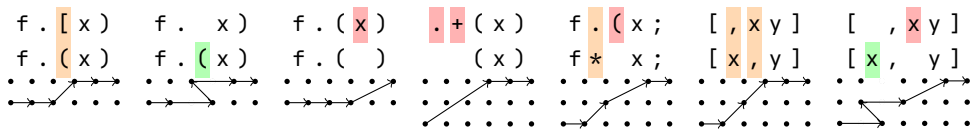
Alternatively, this transition system can be viewed as a kind of proof system in an unlabeled lattice. This is equivalent to Schultz and Mihov's automaton, but more amenable to our purposes as it does not any contain ε-arcs, and uses skip connections to represent consecutive deletions.

$$\frac{s \in \Sigma \quad i \in [0,n] \quad j \in [1,k]}{(q_{i,j-1} \xrightarrow{s} q_{i,j}) \in \delta} \qquad \frac{s \in \Sigma \quad i \in [1,n] \quad j \in [1,k]}{(q_{i-1,j-1} \xrightarrow{s} q_{i,j}) \in \delta}$$

$$\frac{i \in [1,n] \quad j \in [0,k]}{(q_{i-1,j} \xrightarrow{\sigma_i} q_{i,j}) \in \delta} \qquad \frac{d \in [1,d_{\max}] \quad i \in [d+1,n] \quad j \in [d,k]}{(q_{i-d-1,j-d} \xrightarrow{\sigma_i} q_{i,j}) \in \delta}$$

$$\frac{}{q_{0,0} \in I}\ \textsc{Init} \qquad \frac{q_{i,j} \qquad |n-i+j| \le k}{q_{i,j} \in F}\ \textsc{Done}$$

Each arc plays a specific role. ↖ handles insertions, ↗ handles substitutions, ↖ handles insertions and ⋰ handles [consecutive] deletions of various lengths. Let us consider some illustrative cases.



Note that the same patch can have multiple Levenshtein alignments. Done constructs the final states, which are all states accepting strings σ' such that Levenshtein distance of $\Delta(\sigma, \sigma') \le d_{\max}$.

To avoid creating a parallel bundle of arcs for each insertion and substutition point, we instead decorate each arc with a nominal predicate, accepting or rejecting σ_i. To distinguish this nominal variant from the original construction, we highlight the modified rules in orange below.

$$\frac{i \in [0,n] \quad j \in [1,k]}{(q_{i,j-1} \xrightarrow{[\neq \sigma_i]} q_{i,j}) \in \delta} \qquad \frac{i \in [1,n] \quad j \in [1,k]}{(q_{i-1,j-1} \xrightarrow{[\neq \sigma_i]} q_{i,j}) \in \delta}$$

$$\frac{i \in [1,n] \quad j \in [0,k]}{(q_{i-1,j} \xrightarrow{[=\sigma_i]} q_{i,j}) \in \delta} \qquad \frac{d \in [1,d_{\max}] \quad i \in [d+1,n] \quad j \in [d,k]}{(q_{i-d-1,j-d} \xrightarrow{[=\sigma_i]} q_{i,j}) \in \delta}$$

Nominalizing the NFA eliminates the creation of $e = |\Sigma - 1| \cdot |\sigma| \cdot 2d_{\max}$ unnecessary arcs over the entire Levenshtein automaton and drastically reduces the size of the construction to follow, but does not affect the underlying semantics. Thus, it is essential to first nominalize the automaton before proceeding to avoid a large blowup in the intermediate grammar.

4.3 Levenshtein-Bar-Hillel Construction

We now describe the Bar-Hillel construction, which generates a grammar recognizing the intersection between a regular and a context-free language, then specialize it to Levenshtein intersections.

Lemma 4.1. *For any context-free language ℓ and finite state automaton α, there exists a context-free grammar $G_\cap$ such that $\mathcal{L}(G_\cap) = \ell \cap \mathcal{L}(\alpha)$. See Bar-Hillel [7].*

Although Bar-Hillel [7] lacks an explicit construction, Beigel and Gasarch [10] construct $G_\cap$ like so:

$$\frac{q \in I \quad r \in F}{(S \to qSr) \in P_\cap} \qquad \frac{(A \to a) \in P \qquad (q \xrightarrow{a} r) \in \delta}{(qAr \to a) \in P_\cap} \qquad \frac{(w \to xz) \in P \qquad p,q,r \in Q}{(pwr \to (pxq)(qzr)) \in P_\cap} \bowtie$$

This, now standard, Bar-Hillel construction applies to any CFL and REG language intersection, but generates a grammar whose cardinality is approximately $|P_\cap| = |I| \cdot |F| + |P| \cdot |\Sigma| \cdot |\sigma| \cdot 2d_{\max} + |P| \cdot |Q|^3$. Applying the BH construction directly to practical langauges and code snippets can generate hundreds of trillions of productions for even moderately sized grammars and Levenshtein automata. Instead, we will describe a kind of reachability analysis that elides many unreachable productions in the case of Levenshtein intersection.

Consider $\bowtie$, the most expensive rule. What $\bowtie$ tells us is each nonterminal in the intersection grammar, $P_\cap$, matches a substring simultaneously recognized by (1) a pair of states in the original NFA and (2) a nonterminal in the original CFG. A key observation is that $\bowtie$ considers the intersection between every such triple, but this is a gross overapproximation for most NFAs and CFGs, as the vast majority of all state pairs and nonterminals recognize no strings in common.

To identify these triples, we define an interval domain that soundly overapproximates the Parikh image, encoding the minimum and maximum number of terminals each nonterminal can generate. Since some intervals may be right-unbounded, we write $\mathbb{N}^* = \mathbb{N} \cup \{\infty\}$ to denote the upper bound, and $\Pi = \{[a,b] \in \mathbb{N} \times \mathbb{N}^* \mid a \leq b\}^{|\Sigma|}$ to denote the Parikh map of all terminals.

Definition 4.2 (Parikh mapping of a nonterminal). Let $p : \Sigma^* \to \mathbb{N}^{|\Sigma|}$ be the Parikh operator [31], which counts the frequency of terminals in a string. Let $\pi : V \to \Pi$ be a function returning the smallest interval such that $\forall \sigma : \Sigma^*, \forall v : V, v \Rightarrow^* \sigma \vdash p(s) \in \pi(v)$.

In other words, the Parikh mapping computes the greatest lower and least upper bound of the Parikh image over all strings in the language of a nonterminal. The infimum of a nonterminal's Parikh interval tells us how many of each terminal a nonterminal *must* generate, and the supremum tells us how many it *can* generate. Likewise, we define a similar relation over NFA state pairs:

Definition 4.3 (Parikh mapping of NFA states). We define $\pi : Q \times Q \to \Pi$ as returning the smallest interval such that $\forall \sigma : \Sigma^*, \forall q, q' : Q, q \overset{\sigma}{\Longrightarrow} q' \vdash p(\sigma) \in \pi(q, q')$.

Next, we will define a measure on Parikh intervals, which will represent the minimum total edits required to transform a string in one Parikh interval to a string in another.

Definition 4.4 (Parikh divergence). Given two Parikh intervals $\pi, \pi' : \Pi$, we define the divergence between them as $\pi \parallel \pi' = \sum_{n=1}^{|\Sigma|} \min_{(i,i') \in \pi[n] \times \pi'[n]} |i - i'|$.

Now, we know that if the Parikh divergence between two intervals exceeds the Levenshtein margin between two states in a Lev-NFA, those intervals must be incompatible as no two strings, one from each Parikh interval, can be transformed into the other with fewer than $\pi \parallel \pi'$ edits.

Definition 4.5 (Levenshtein-Parikh compatibility). Let $q = q_{h,i}, q' = q_{j,k}$ be two states in a Lev-NFA and V be a CFG nonterminal. We say that $(q, v, q') : Q \times V \times Q$ are compatible iff the Parikh divergence is bounded by the Levenshtein margin $k - i$, i.e., $v \lhd qq' \iff (\pi(v) \parallel \pi(q, q')) \leq k - i$.

We define the modified Bar-Hillel construction for Levenshtein intersections as follows:

$$\frac{(A \to a) \in P \quad S.a \quad (q \xrightarrow{S} r) \in \delta}{(qAr \to a) \in P_\cap} \qquad \frac{w \lhd pr \quad x \lhd pq \quad z \lhd qr \quad (w \to xz) \in P \quad p,q,r \in Q}{(pwr \to (pxq)(qzr)) \in P_\cap} \bowtie$$

Finally, once $G_\cap$ is constructed, it is renormalized by removing all unreachable and non-generating productions following [22] to obtain $G^*_\cap$, which is usually several orders of magnitude smaller. When this grammar is sufficiently small, we can extract all relevant repairs from it, otherwise we sample from it to obtain a subset of candidate repairs.

4.4 Repair as idempotent matrix completion

Now that we have a language that recognizes local repairs, we need a method to extract the repairs themselves. We impose specific criteria on such a procedure: it must generate (1) only valid repairs and (2) all repairs in the langauge. We also require the sampler to have (3) parallel decomposability and generate samples (4) uniformly, both (5) with and (6) without replacement in (7) constant time and space. These requirements motivate the need for a specialized generating function.

In this section, we will introduce the porous completion problem and show how it can be translated to a kind of idempotent matrix completion, whose roots are valid strings in a context-free language. This technique is convenient for its geometric interpretability, amenability to parallelization, and generalizability to any CFG, regardless of finitude or ambiguity. We will see how, by redefining the algebraic operations $\oplus, \otimes$ over different carrier sets, one can obtain a recognizer, parser, generator, Parikh map and other convenient structures for working with CFLs.

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

$$X \otimes Z \coloneqq \left\{ w \mid \langle x, z\rangle \in X \times Z, (w \to xz) \in P \right\} \tag{2}$$

If we define $\hat{\sigma}_r \coloneqq \{w \mid (w \to \sigma_r) \in P\}$, then construct a matrix with nonterminals on the superdiagonal representing each token, $M_0[r+1=c](G', \sigma) \coloneqq \hat{\sigma}_r$, the fixpoint $M_{i+1} = M_i + M_i^2$ is uniquely determined by the superdiagonal entries, which

$$M_0 \coloneqq \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \varnothing & \cdots & \varnothing \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \varnothing \\ & & & \ddots & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \Lambda & \cdots & \varnothing \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \Lambda \\ & & & \ddots & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \ldots \Rightarrow M_\infty = \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \Lambda & \cdots & \Lambda^*_\sigma \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \Lambda \\ & & & \ddots & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix}$$

The northeasternmost entry Λ^*_σ gives us the recognizer $R(G', \sigma) \coloneqq [S \in \Lambda^*_\sigma] \Leftrightarrow [\sigma \in \mathcal{L}(G)]$ [1].

This procedure can be lifted to the domain of strings containing free variables, which we call the *porous completion problem*, whose solutions are the set of all syntactically valid strings consistent with the input string.

Definition 4.6 (Porous completion). Let $\underline{\Sigma} \coloneqq \Sigma \cup \{\_\}$, where _ denotes a hole. We denote $\sqsubseteq : \Sigma^n \times \underline{\Sigma}^n$ as the relation $\{\langle \sigma', \sigma\rangle \mid \sigma_i \in \Sigma \implies \sigma'_i = \sigma_i\}$ and the set of all inhabitants $\{\sigma' : \Sigma^+ \mid \sigma' \sqsubseteq \sigma\}$ as $\mathrm{H}(\sigma)$. Given a *porous string*, $\sigma : \underline{\Sigma}^*$ we seek all syntactically valid inhabitants, i.e., $A(\sigma) \coloneqq \mathrm{H}(\sigma) \cap \ell$.

Let us consider an example with two holes, $\sigma = 1 \_ \_$, and the grammar being $G \coloneqq \{S \to NON, O \to + \mid \times, N \to 0 \mid 1\}$. This can be rewritten into CNF as $G' \coloneqq \{S \to NL, N \to 0 \mid 1, O \to \times \mid +, L \to ON\}$. Using the algebra where $\oplus = \cup$, $X \otimes Z = \left\{ w \mid \langle x, z\rangle \in X \times Z, (w \to xz) \in P \right\}$, the fixpoint $M' = M + M^2$ can be computed as follows, shown in the leftmost column:

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

	2^V	$\mathbb{Z}_2^{\lvert V\rvert}$	$\mathbb{Z}_2^{\lvert V\rvert} \to \mathbb{Z}_2^{\lvert V\rvert}$
M_0	$\begin{pmatrix} \{N\} & & \\ & \{N,O\} & \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} \square\blacksquare\square\square & & \\ & \square\blacksquare\blacksquare\square & \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & & \\ & V_{1,2} & \\ & & V_{2,3} \end{pmatrix}$
M_1	$\begin{pmatrix} \{N\} & \varnothing & \\ & \{N,O\} & \{L\} \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} \square\blacksquare\square\square & \square\square\square\square & \\ & \square\blacksquare\blacksquare\square & \blacksquare\square\square\square \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & V_{0,2} & \\ & V_{1,2} & V_{1,3} \\ & & V_{2,3} \end{pmatrix}$
M_∞	$\begin{pmatrix} \{N\} & \varnothing & \{S\} \\ & \{N,O\} & \{L\} \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} \square\blacksquare\square\square & \square\square\square\square & \square\square\square\blacksquare \\ & \square\blacksquare\blacksquare\square & \blacksquare\square\square\square \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & V_{0,2} & V_{0,3} \\ & V_{1,2} & V_{1,3} \\ & & V_{2,3} \end{pmatrix}$

The same procedure can be translated, without loss of generality, into the bit domain ($\mathbb{Z}_2^{|V|}$) using a lexicographic ordering, however M_∞ in both 2^V and $\mathbb{Z}_2^{|V|}$ represents a decision procedure, i.e., $[S \in V_{0,3}] \Leftrightarrow [V_{0,3,3} = \blacksquare] \Leftrightarrow [A(\sigma) \neq \varnothing]$. Since $V_{0,3} = \{S\}$, we know there exists at least one $\sigma' \in A$, but M_∞ does not reveal its identity.

In order to extract the inhabitants, we can translate the bitwise procedure into an equation with free variables. Here, we can encode the idempotency constraint directly as $M = M^2$. We first define $X \boxtimes Z = [X_2 \wedge Z_1, \bot, \bot, X_1 \wedge Z_0]$ and $X \boxplus Z = [X_i \vee Z_i]_{i \in [0,|V|)}$. Since the unit nonterminals O, N can only occur on the superdiagonal, they may be safely ignored by $\boxtimes$. To solve for M_∞, we proceed by first computing $V_{0,2}, V_{1,3}$ as follows:

$$\begin{aligned} V_{0,2} &= V_{0,j} \cdot V_{j,2} = V_{0,1} \boxtimes V_{1,2} \\ &= [L \in V_{0,2}, \bot, \bot, S \in V_{0,2}] \\ &= [O \in V_{0,1} \wedge N \in V_{1,2}, \bot, \bot, N \in V_{0,1} \wedge L \in V_{1,2}] \\ &= [V_{0,1,2} \wedge V_{1,2,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,2,0}] \end{aligned}$$

$$\begin{aligned} V_{1,3} &= V_{1,j} \cdot V_{j,3} = V_{1,2} \boxtimes V_{2,3} \\ &= [L \in V_{1,3}, \bot, \bot, S \in V_{1,3}] \\ &= [O \in V_{1,2} \wedge N \in V_{2,3}, \bot, \bot, N \in V_{1,2} \wedge L \in V_{2,3}] \\ &= [V_{1,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{1,2,1} \wedge V_{2,3,0}] \end{aligned}$$

Now we solve for the corner entry $V_{0,3}$ by taking the bitwise dot product between the first row and last column, yielding:

$$\begin{aligned} V_{0,3} &= V_{0,j} \cdot V_{j,3} = V_{0,1} \boxtimes V_{1,3} \boxplus V_{0,2} \boxtimes V_{2,3} \\ &= [V_{0,1,2} \wedge V_{1,3,1} \vee V_{0,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0}] \end{aligned}$$

Since we only care about $V_{0,3,3} \Leftrightarrow [S \in V_{0,3}]$, so we can ignore the first three entries and solve for:

$$\begin{aligned} V_{0,3,3} &= V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0} \\ &= V_{0,1,1} \wedge (V_{1,2,2} \wedge V_{2,3,1}) \vee V_{0,2,1} \wedge \bot \\ &= V_{0,1,1} \wedge V_{1,2,2} \wedge V_{2,3,1} \\ &= [N \in V_{0,1}] \wedge [O \in V_{1,2}] \wedge [N \in V_{2,3}] \end{aligned}$$

Now we know that $\sigma = 1\,\underline{O}\,\underline{N}$ is a valid solution, and we can take the product $\{1\} \times \hat{\sigma}_2^{-1}(O) \times \hat{\sigma}_3^{-1}(N)$ to recover the admissible set, yielding $A = \{1{+}0, 1{+}1, 1{\times}0, 1{\times}1\}$. In this case, since G is unambiguous, there is only one parse tree satisfying $V_{0,|\sigma|,3}$, but in general, there can be multiple valid parse trees.

4.5 An algebraic datatype for context-free parse forests

This procedure generates solutions satisfying the matrix fixpoint, but forgets their provenance. The question naturally arises, is there a way to solve for the parse trees directly? This would allow us to handle ambiguous grammars, whilst preserving the natural structure of the parsing domain.

We will now describe a datatype for compactly representing CFL parse forests, then redefine the semiring algebra over this domain. This construction is especially convenient for tracking provenance under ambiguity, Parikh mapping, counting the size of a finite CFL, and sampling trees either with replacement using a PCFG, or without replacement using an integer pairing function.

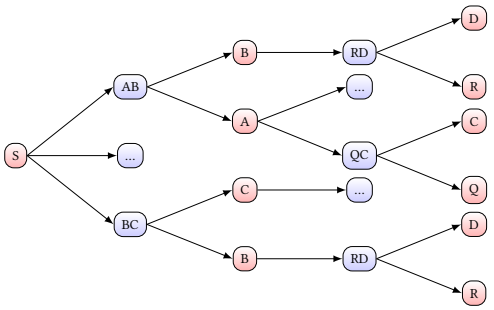


Fig. 3. A partial $\mathbb{T}_2$ corresponding to the grammar $\{S \to BC \mid \ldots \mid AB, B \to RD \mid \ldots, A \to QC \mid \ldots\}$.

We first define a datatype $\mathbb{T}_3 = (V \cup \Sigma) \rightharpoonup \mathbb{T}_2$ where $\mathbb{T}_2 = (V \cup \Sigma) \times (\mathbb{N} \rightharpoonup \mathbb{T}_2 \times \mathbb{T}_2)$[2]. Morally, we can think of $\mathbb{T}_2$ as an implicit set of possible trees that can be generated by a CFG in CNF, consistent with a finite-length porous string. Structurally, we may interpret $\mathbb{T}_2$ as an algebraic data type corresponding to the fixpoints of the following recurrence. This tells us each $\mathbb{T}_2$ can be a terminal, nonterminal, or a nonterminal and a sequence of nonterminal pairs and their two children:

$$L(p) = 1 + pL(p) \qquad P(a) = \Sigma + V + VL\left(V^2 P(a)^2\right) \tag{3}$$

Given a $\sigma : \underline{\Sigma}$, we construct $\mathbb{T}_2$ from the bottom-up, and sample from the top-down. Depicted in Fig. 3 is a partial $\mathbb{T}_2$, where red nodes are `roots` and blue nodes are `children`.

We construct the first upper diagonal $\hat{\sigma}_r = \Lambda(\sigma_r)$ as follows:

$$\Lambda(s : \underline{\Sigma}) \mapsto \begin{cases} \bigoplus_{s' \in \Sigma} \Lambda(s') & \text{if } s \text{ is a hole,} \\ \left\{\mathbb{T}_2\left(w, \left[\langle \mathbb{T}_2(s), \mathbb{T}_2(\varepsilon)\rangle\right]\right) \mid (w \to s) \in P\right\} & \text{otherwise.} \end{cases} \tag{4}$$

This initializes the superdiagonal entries, enabling us to compute the fixpoint M_∞ by redefining $\oplus, \otimes : \mathbb{T}_3 \times \mathbb{T}_3 \to \mathbb{T}_3$ as:

$$X \oplus Z \mapsto \bigcup_{k \in \pi_1(X \cup Z)} \left\{k \Rightarrow \mathbb{T}_2(k, x \cup z) \mid x \in \pi_2(X \circ k), z \in \pi_2(Z \circ k)\right\} \tag{5}$$

$$X \otimes Z \mapsto \bigoplus_{(w \to xz) \in P} \left\{\mathbb{T}_2\Big(w, \left[\langle X \circ x, Z \circ z\rangle\right]\Big) \mid x \in \pi_1(X), z \in \pi_1(Z)\right\} \tag{6}$$

These operators group subtrees by their root nonterminal, then aggregate their children. Instead of tracking sets, each Λ now becomes a dictionary of $\mathbb{T}_2$, indexed by their root nonterminals.

$\mathbb{T}_2$ is a convenient datatype for many operations involving CFGs. We can use it to approximate the Parikh image, compute the size of a finite CFG, and sample parse trees with or without replacement. For example, to obtain the Parikh map, we may use the following recurrence,

$$\pi(T : \mathbb{T}_2) \mapsto \begin{cases} \left[[1,1] \text{ if } \texttt{root}(T) = s \text{ else } [0,0]\right]_{s \in \Sigma} & \text{if } T \text{ is a leaf,} \\ \bigoplus_{\langle T_1, T_2\rangle \in \texttt{children}(T)} \pi(T_1) \otimes \pi(T_2) & \text{otherwise.} \end{cases} \tag{7}$$

where the operations over Parikh maps $\oplus, \otimes : \Pi \times \Pi \to \Pi$ are defined respectively as follows:

[2]Given a $T : \mathbb{T}_2$, we may also refer to $\pi_1(T), \pi_2(T)$ as $\texttt{root}(T)$ and $\texttt{children}(T)$ respectively.

$$X \oplus Z \mapsto \left[\left[\min(X_s \cup Z_s), \max(X_s \cup Z_s)\right]\right]_{s \in \Sigma} \tag{8}$$

$$X \otimes Z \mapsto \left[\left[\min(X_s) + \min(Z_s), \max(X_s) + \max(Z_s)\right]\right]_{s \in \Sigma} \tag{9}$$

To obtain the Parikh interval for a CFG up to finite length bounds, we abstractly parse the string $(\_)^n$ for all $n \in [1, l_{\max}]$ and take the union of the intervals across $[|\underset{\sim}{\sigma}| - d_{\max}, |\underset{\sim}{\sigma}| + d_{\max}]$, which subsumes every valid repair in the Levenshtein ball.

$\mathbb{T}_2$ also allows us to sample whole parse trees by obtaining $(\Lambda^*_\sigma \circ S) : \mathbb{T}_2$. Given a probabilistic CFG whose productions indexed by each nonterminal are decorated with a probability vector $\mathbf{p}$ (this may be uniform in the non-probabilistic case), we define a tree sampler $\Gamma : (\mathbb{T}_2 \mid \mathbb{T}_2^2) \rightsquigarrow \mathbb{T}$ which recursively samples children according to a Multinoulli distribution:

$$\Gamma(T) \mapsto \begin{cases} \texttt{BTree}\Big(\texttt{root}(T), \Gamma\big(\text{Multi}(\texttt{children}(T), \mathbf{p})\big)\Big) & \text{if } T : \mathbb{T}_2 \\ \Big\langle \Gamma\big(\pi_1(T)\big), \Gamma\big(\pi_2(T)\big) \Big\rangle & \text{if } T : \mathbb{T}_2 \times \mathbb{T}_2 \end{cases} \tag{10}$$

This is closely related to the generating function for the ordinary Boltzmann sampler from analytic combinatorics,

$$\Gamma C(x) \mapsto \begin{cases} \text{Bern}\left(\frac{A(x)}{A(x)+B(x)}\right) \to \Gamma A(x) \mid \Gamma B(x) & \text{if } \mathcal{C} = \mathcal{A} + \mathcal{B} \\ \big\langle \Gamma A(x), \Gamma B(x) \big\rangle & \text{if } \mathcal{C} = \mathcal{A} \times \mathcal{B} \end{cases} \tag{11}$$

however unlike Duchon et al. [20], our work does not depend on rejection to guarantee exact-size sampling, as all trees contained in $\mathbb{T}_2$ will necessarily be the same width.

The number of binary trees inhabiting a single instance of $\mathbb{T}_2$ is sensititive to the number of nonterminals and rule expansions in the grammar. To obtain the total number of trees with breadth n, we abstractly parse the porous string $\sigma = (\_)^n$, letting $T = \Lambda^*_\sigma \circ S$, then use the recurrence below to compute the total number of trees:

$$|T : \mathbb{T}_2| \mapsto \begin{cases} 1 & \text{if } T \text{ is a leaf,} \\ \sum_{\langle T_1, T_2 \rangle \in \texttt{children}(T)} |T_1| \cdot |T_2| & \text{otherwise.} \end{cases} \tag{12}$$

To sample all trees in a given $T : \mathbb{T}_2$ uniformly without replacement, we then construct a modular pairing function $\varphi : \mathbb{T}_2 \to \mathbb{Z}_{|T|} \to \texttt{BTree}$, that we define as follows:

$$\varphi(T : \mathbb{T}_2, i : \mathbb{Z}_{|T|}) \mapsto \begin{cases} \Big\langle \texttt{BTree}\big(\texttt{root}(T)\big), i \Big\rangle & \text{if } T \text{ is a leaf,} \\ \text{Let } b = |\texttt{children}(T)|, & \\ \quad q_1, r = \Big\langle \lfloor \frac{i}{b} \rfloor, i \pmod b \Big\rangle, & \\ \quad lb, rb = \texttt{children}[r], & \\ \quad T_1, q_2 = \varphi(lb, q_1), & \\ \quad T_2, q_3 = \varphi(rb, q_2) \text{ in} & \\ \Big\langle \texttt{BTree}\big(\texttt{root}(T), T_1, T_2\big), q_3 \Big\rangle & \text{otherwise.} \end{cases} \tag{13}$$

If the language is suffiently small, instead of sampling trees, we can sample integers uniformly without replacement from $\mathbb{Z}_{|T|}$ then decode them into trees using φ. This procedure is the basis for our sampling algorithm and the method we use to decode repairs from the intersection grammar.

4.6 Ranked Repair

Returning to the ranked repair problem (Def. 3.2), the above procedure returns a set of strings, and we need an ordering over them. We note that any metric is sufficient, such as the perplexity of the string under a large language model, or the probability of the string under a PCFG. We the simplest possible solution: the negative log likelihood of the string computed by a variable-order Markov chain. This is a simple, fast, and effective metric for ranking strings, and as we will show, already yields competitive results in practice.

5 EVALUATION

We consider the following research questions for our evaluation:

- **RQ 1**: What properties do natural repairs exhibit? (e.g., error frequency, edit distance)
- **RQ 2**: How precise is our technique at fixing syntax errors? (Precision@k vs. Seq2Parse)
- **RQ 3**: Which design choices are most significant? (e.g., search vs. sampling, n-gram order)

5.1 Dataset

For our evaluation, we use the StackOverflow dataset from [24]. We preprocess the dataset to lexicalize both the broken and fixed code snippets, then filter the dataset by length and edit distance, in which all Python snippets whose broken form is fewer than 80 lexical tokens and whose human fix is under four Levenshtein edits is retained.

In the following experiments, we use two datasets of Python snippets. The first is a set of 5,600 pairwise-aligned (broken, fixed) Python code snippets from Wong et al.'s StackOverflow dataset [37] shorter than 40 lexical tokens, whose patch sizes are less than five lexical tokens ($|\Sigma| = 50, |\underset{\sim}{\sigma}| \leq 40, \Delta(\underset{\sim}{\sigma}, \ell) \leq 4$).

The StackOverflow dataset is comprised of 500k Python code snippets, each of which has been annotated with a human repair. We depict the normalized edit loations relative to the snippet length below.

In the second set of experiments, we use a dataset of valid Python snippets from BIFI [38], then synthetically corrupt them by introducing syntax errors sampled from a distribution trained from the StackOverflow dataset. We then measure the precision@k of our repair procedure at recovering the original, uncorrupted snippet.

5.2 Experimental Setup

We measure the precision@k of our repair procedure at recovering human repairs of varying edit distances and latency cutoffs, where both the broken and fixed code are written by a human.

To train the scoring function, we use a length-5 variable-order Markov (VOM) chain implemented using a count-min sketch based on Apache Datasketches [5]. Training on 55 million StackOverflow tokens took roughly 10 minutes, after which calculating perplexity is nearly instantaneous. Sequences are scored using negative log likelihood with Laplace smoothing and our evaluation measures the precision@{1, 5, 10, All} for samples at varying latency cutoffs.

Both sets of experiments were conducted on 40 Intel Skylake cores running at 2.4 GHz, with 16 GB of RAM, running bytecode compiled for JVM 17.0.2. We measure the precision using abstract lexical matching, following the Seq2Parse [33] evaluation, and give a baseline for their approach on the same dataset.

5.3 StackOverflow Evaluation

For our first experiment, we run the sampler until the human repair is detected, then measure the number of samples required to draw the exact human repair across varying Levenshtein radii.

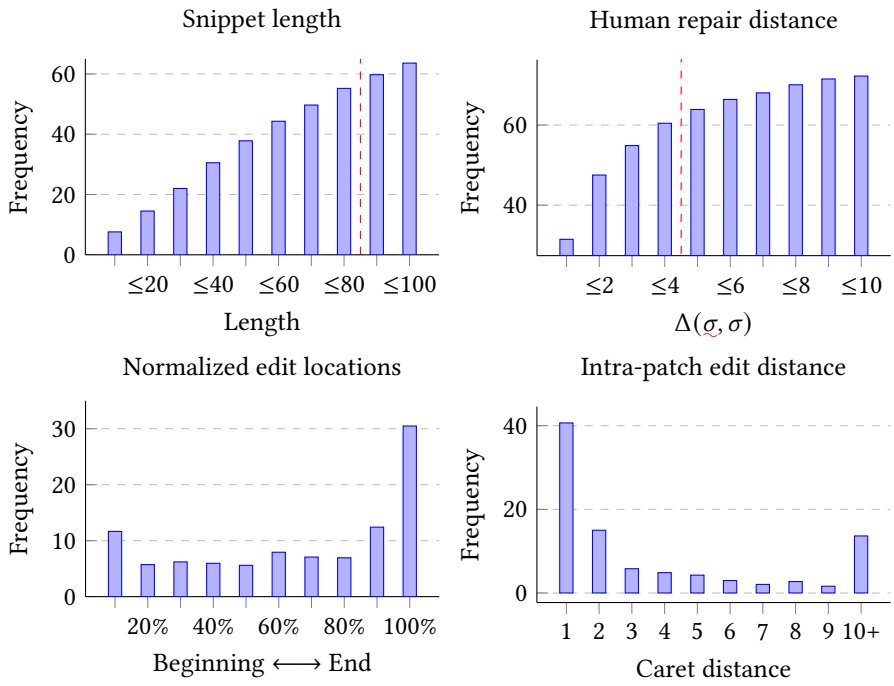


Fig. 4. Repair statistics across the StackOverflow dataset, of which our method can handle about half in ~30s and ~140GB. Larger repairs and Levenshtein radii are possible, albeit requiring additional time and memory.

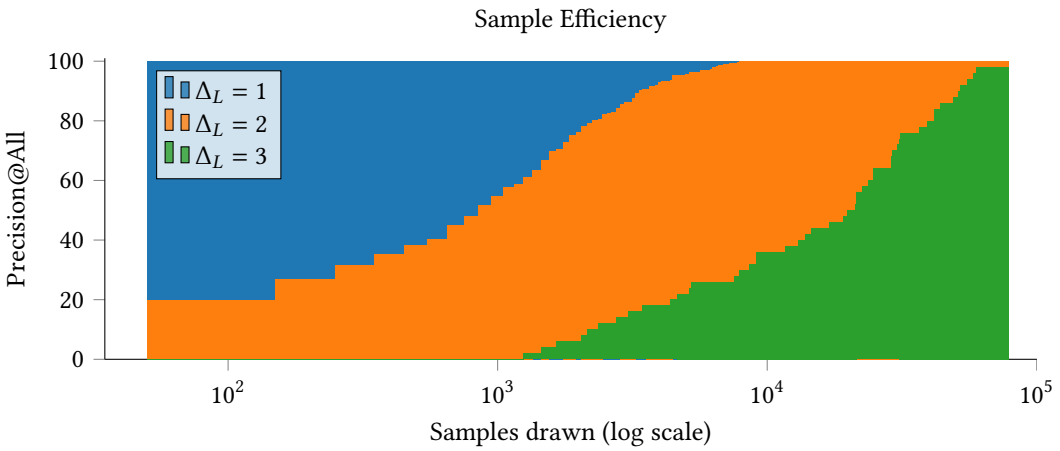


Fig. 5. Sample efficiency of LBH sampler at varying Levenshtein radii. After drawing up to $\sim 10^5$ samples without replacement we can usually saturate the admissible set for almost all repairs fewer than four edits.

Next, measure the precision at various ranking cutoffs for varying wall-clock timeouts. Here, P@{k=1, 5, 10, All} indicates the percentage of syntax errors with a human repair of $\Delta = \{1, 2, 3, 4\}$ edits found in $\leq p$ seconds that were matched within the top-k results, based on n-gram likelihood.

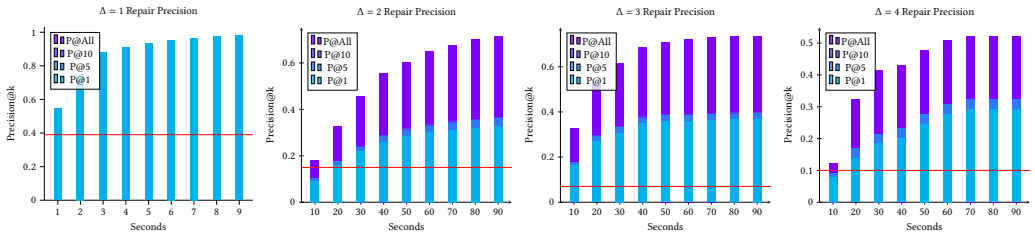


Fig. 6. Human repair benchmark. Note the y-axis across different edit distance plots has varying ranges.

Previously, we used a rejection-based sampler, which did not sample directly from the admissible set, but the entire Levenshtein ball, then rejected invalid samples. Although rejection sampling has a much lower minimum latency threshold, the average time required to attain a fixed precision is much higher. We present the results from the earlier evaluation for comparison below.

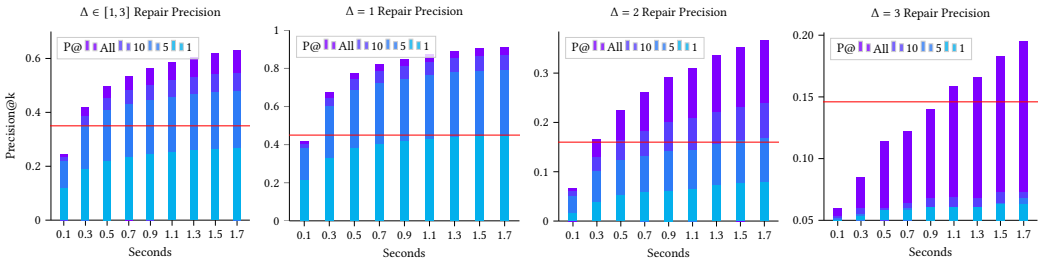


Fig. 7. Adaptive sampling repairs. The red line indicates Seq2Parse precision@1 on the same dataset. Since it only supports generating one repair, we do not report precision@k or the intermediate latency cutoffs.

We also evaluate Seq2Parse on the same dataset. Seq2Parse only supports precision@1 repairs, and so we only report Seq2parse 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 repairs matching the human repair for both our method and Seq2Parse. Seq2Parse latency varies depending on the length of the repair, averaging 1.5s for $\Delta = 1$ to 2.7s for $\Delta = 3$, across the entire StackOverflow dataset.

End-to-end throughput varies significantly with the edit distance of the repair. Some errors are trivial to fix, while others require a large number of edits to be sampled before a syntactically valid edit is discovered. We evaluate throughput by sampling edits across invalid strings $|\underset{\sim}{\sigma}| \leq 40$ from the StackOverflow dataset of varying length, and measure the total number of syntactically valid edits discovered, as a function of string length and language edit distance $\Delta \in [1, 3]$. Each trial is terminated after 10 seconds, and the experiment is repeated across 7.3k total repairs. Note the y-axis is log-scaled, as the number of admissible repairs increases sharply with language edit distance. Our approach discovers a large number of syntactically valid repairs in a relatively short amount of time, and is able to quickly saturate the admissible set for $\Delta(\underset{\sim}{\sigma}, \sigma) \in [1, 4]$ before timeout. As Seq2Parse only generates one syntactically valid repair per string, we do not report its throughput.

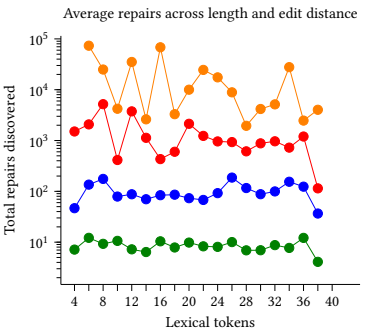


Fig. 8. Total repairs found in 30s.

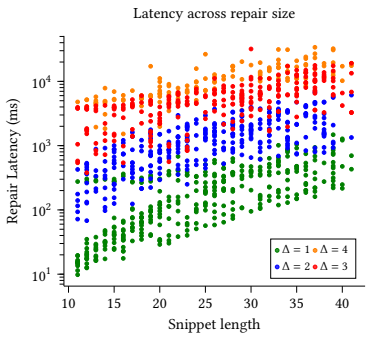


Fig. 9. End-to-end repair timings across snippet length.

In Fig. 9, we plot the end-to-end repair timings across snippet length. To measure the timings, we collect 1000 repair samples of varying lengths and edit distances, then measure the time until the sampler locates the exact human repair. For each code snippet, we measure the total time taken, for repairs between 1-4 Levenshtein edits. Note the logarithmic scale on the y-axis.

5.4 Synthetic repair benchmark

Pairwise naturally-occurring errors and human fixes are the most authentic source of real-world syntax repairs, but can be difficult to obtain due to the paucity of parallel syntax error corpi. In the absence of natural syntax repairs, one viable alternative is to collect a dataset of syntactically valid code, and synthetically corrupt it. The original source code becomes the ground truth repair for the synthetically generated typo, and the target for evaluating the precision of our repair procedure.

In our synthetic repair benchmark, we synthetically corrupt BIFI code snippets using a learned corruption model. To train the model we compute the Levenshtein alignment on the StackOverflow dataset, then approximate the conditional probability of each edit given the local context. During evaluation, we sample a corruption from the learned typo distribution, and measure the precision of our model at recovering the originally valid code sequence.

Suppose we have a dataset of single token edits and their local context. For simplicity, we shall assume a trigram language model, i.e., $P(\sigma'_i \mid \sigma_{i-1}, \sigma_i, \sigma_{i+1})$, however the approach can be generalized to higher-order Markov models. Given a string σ, we can sample error trajectories $q^1(\sigma), q^2(\sigma), \ldots, q^n(\sigma)$ by defining $q(\sigma)$ to sample a single edit from the set of all relevant edit actions $Q(\sigma)$, then recursively applying q to the resulting string. More formally,

(1) Given a string σ, compute $Q(\sigma)$, the set of all relevant edit actions for all possible edit locations by unioning the set of all possible edits at each location, i.e., $Q(\sigma) \coloneqq \bigcup_{i=1}^{|\sigma|-1} \{\sigma'_i \mid 0 < P(\sigma_i \mid \sigma_{i-1}, \sigma_i, \sigma_{i+1})\}$.
(2) Renormalize the probabilities of each edit $P(q \mid \sigma)$ by $\sum_{q \in Q(\sigma)} P(q)$. This ensures the probability of sampling a particular edit is proportional to its relative probability under the language model and sums to 1.
(3) Sample an edit $q(\sigma) \sim Q(\sigma)$, then repeat for n steps where n is sampled from a geometric distribution with mean μ matching the average edit distance of the dataset (this assumes the edit distance is independent of the edits).

This allows us to sample a corruption of a good string matching the distribution of typos in the dataset. We then measure the precision at recovering the originally valid string.

In addition to the StackOverflow dataset, we also evaluate our approach on two datasets containing synthetic strings generated by a Dyck language, and bracketing errors of synthetic and organic provenance in organic source code. The first dataset contains length-50 strings sampled from various Dyck languages, i.e., the Dyck language containing n different types of balanced parentheses. The second contains abstracted Java and Python source code mined from GitHub repositories. The Dyck languages used in the remaining experiments are defined by the following context-free grammar(s):

```
Dyck-1 -> ( ) | ( Dyck-1 ) | Dyck-1 Dyck-1
```

34

34

```
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 experiment (1a), we sample a random valid string $\sigma \sim \Sigma^{50} \cap \mathcal{L}_{\text{Dyck-n}}$, then replace a fixed number of indices in $[0, |\sigma|)$ with holes and measure the average time required to decode ten syntactically-admissible repairs across 100 trial runs. In experiment (1b), we sample a random valid string as before, but delete p tokens at random and rather than provide their 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.

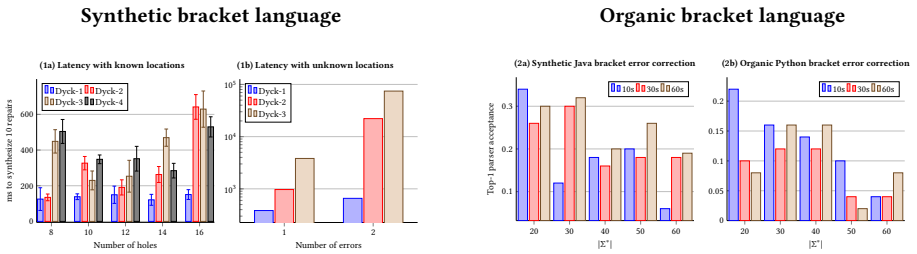


Fig. 10. Benchmarking bracket correction latency and accuracy across two bracketing languages, one generated from Dyck-n, and the second uses an abstracted source code snippet with imbalanced parentheses.

In the second set of experiments, we analyze bracketing errors in a dataset of Java and Python code snippets mined from open-source repositories on GitHub using the Dyck-nw[3], in which all source code tokens except brackets are replaced with a `w` token. For Java (2a), we sample valid single-line statements with bracket nesting more than two levels deep, synthetically delete one bracket uniformly at random, and repair using Tidyparse, then take the top-1 repair after t seconds, and validate using ANTLR's Java 8 parser. For Python (2b), we sample invalid code fragments uniformly from the imbalanced bracket category of the Break-It-Fix-It (BIFI) dataset [38], a dataset of organic Python errors, which we repair using Tidyparse, take the top-1 repair after t seconds, and validate repairs using Python's `ast.parse()` method. Since the Java and Python datasets do not have a ground-truth human fix, we report the percentage of repairs that are accepted by the language's official parser for repairs generated under a fixed time cutoff. Although the Java and Python datasets are not directly comparable, we observe that Tidyparse can detect and repair a significant fraction of bracket errors in both languages with a relatively unsophisticated grammar.

6 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.

[3]Using the Dyck-n grammar augmented with a single additional production, `Dyck-1` → `w` | `Dyck-1`. Contiguous non-bracket characters are substituted with a single placeholder token, `w`, and restored verbatim after bracket repair.

6.1 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 can be formulated as a dynamic programming problem, as shown by Cocke-Younger-Kasami (CYK) [32], Earley [21] and others. These parsers have roughly cubic complexity with respect to the length of the input string.

As shown by Valiant [36], Lee [27] 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 reduction unlocks the door to a wide range of complexity-theoretic and practical speedups to CFL recognition and fast general parsing algorithms.

Okhotin (2001) [30] extends CFGs with language conjunction in *conjunctive grammars*, followed by Zhang & Su (2017) [39] who apply conjunctive language reachability to dataflow analysis.

6.2 Language equations

Language equations are a powerful tool for reasoning about formal languages and their inhabitants. First proposed by Ginsburg et al. [23] 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 [12, 13] and Antimirov [4], one can take the derivative of a language equation, yielding another equation. This can be interpreted as a kind of continuation or language quotient, revealing the suffixes that complete a given prefix. This technique leads to an elegant family of algorithms for incremental parsing [1, 29] and automata minimization [11]. 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 [29], to our knowledge they have never previously been considered for the purpose of code completion or syntax error correction.

6.3 Syntax repair

In finite languages, syntax repair corresponds to spelling correction, a more restrictive and largely solved problem. Schulz and Stoyan [34] 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 Irons [25] and Aho [2] propose a dynamic programming algorithm to compute the minimum number of edits required to fix an invalid string. Prior work on error correcting parsing only considers the shortest edit(s), and does not study multiple edits over the Levenshtein ball. Furthermore, the problem of actually generating the repairs is not well-posed, as there are usually 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.

6.4 Classical program synthesis

There is related work on string constraint solving in the constraint programing literature, featuring solvers like CFGAnalyzer and HAMPI [26], which consider bounded context free grammars and intersections thereof. Axelson et al. (2008) [6] has some work on incremental SAT encoding but does not exploit the linear-algebraic structure of parsing, conjunctive reachability nor provide real-time guarantees. D'Antoni et al. (2014) introduces *symbolic automata* [17], a generalization of finite automata which 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 context-free and conjunctive grammars with finite alphabets and does not require any special grammar encoding. The matrix encoding makes it particularly amenable to parallelization.

6.5 Error correcting codes

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 scheme 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 [18, 35] or Levenshtein [8, 9] balls, from which we draw inspiration for our work.

6.6 Neural program repair

The recent success of deep learning has lead to a variety of work on neural program repair [3, 15, 19]. These approaches typically employ Transformer-based neural language models (NLMs) and model the problem as a sequence-to-sequence transformation. Although recent work on circuit lower bounds have cast doubt on the ability of Transformers to truly learn formal languages [14, 28], expressivity aside, these models have been widely adopted for practical program repair tasks. In particular, two papers stand out being most closely related to our own: Break-It-Fix-It (BIFI) [38] and Seq2Parse [33]. BIFI adapts techniques from semi-supervised machine translation to generate synthetic errors in clean code and fixes them. This reduces the amount of pairwise training data, but may generalize poorly to natural errors. Seq2Parse combines a transformer-based model with an augmented version of the Early parser to suggest error rules, but only suggests a single repair. Our work differs from both in that we suggest multiple repairs with much lower latency, do not require a pairwise repair dataset, and can fix syntax errors in any language with a well-defined grammar. We note our approach is complementary to existing work in neural program repair, and may be used to generate synthetic repairs for training or employ a NLM model to rank its repairs.

7 DISCUSSION

The main lesson we draw from our experiments is that it is possible to leverage compute to compete with large language models on practical program repair tasks. Though extremely sample-efficient, their size comes at the cost of higher latency, and domain adaptation requires fine-tuning or

retraining on pairwise repairs. Our approach uses a tiny grammar and a relatively cheap ranking metric to achieve comparable accuracy at the same latency. This allows us to repair errors in languages with little to no training data and provides far more flexibility and controllability.

Our primary insight leading to SoTA precision@{5,10} is that repairs are typically concentrated near a small number of edit locations, and by biasing the search toward previously successful edit locations, we can achieve a significant speedup over a naïve search. We note this heuristic may not be applicable to all grammars, and it may be possible to construct less natural counterexamples where this heuristic fails, although we consider these unlikely in practice.

Latency can vary depending on many factors including string length and grammar size, and critically the Levenshtein edit distance. This can be an advantage because in the absence of any contextual or statistical information, syntax and Levenshtein edits are often sufficiently constrained to determine a small number of valid repairs. It is also a limitation because as the number of edits grows, the admissible set grows rapidly and the number of valid repairs may become too large to be useful without a good metric, depending on the language and source code snippet under repair.

Tidyparse in its current form has several other technical shortcomings: firstly, it does not incorporate any neural language modeling technology at present, an omission we hope to address. Training a language model to predict likely repair locations and rank admissible results could lead to lower overall latency and more natural repairs. We also hope to explore the use of Metropolis-Hastings and determinantal point processes to encourage sampling diversity.

Secondly, our current method does not specialize language intersection to the grammar family, nor employ Bar-Hillel's [7] construction for REG-CFL intersection, which would lead to a more efficient encoding of Levenshtein-CFL reachability. Furthermore, considering recent extensions of Boolean matrix-based parsing to linear context-free rewriting systems (LCFRS) [16], it may be feasible to search through richer language families within the SAT solver without employing an external stochastic search to generate and validate candidate repairs.

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. This can be partly alleviated by filtering the results through an incremental compiler or linter, however, the latency introduced may be non-negligible. It is also possible to encode type-based semantic constraints into the solver and we intend to explore this direction more fully in future work.

We envision a few primary use cases for our tool: (1) helping novice programmers become more quickly familiar with a new programming language, (2) autocorrecting common typos among proficient but forgetful programmers, (3) as a prototyping tool for PL designers and educators, and (4) as a pluggable library or service for parser-generators and language servers. 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 benchmarking and testing.

8 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 has focused on general λ-calculi, or narrow languages such as finite sets or regular expressions. The former are too expressive to be efficiently 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 communications, 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.

Over the last few years, there has been a surge of progress in applying language models to write programs. That work is primarily based on methods from differential calculus and continuous optimization, leading to the so-called *naturalness hypothesis*, which suggests programming languages are not so different from natural ones. In contrast, programming language theory takes the view that languages are essentially discrete and finitely-generated sets, although perhaps uncomputably large ones, governed by logical calculi. Programming, thus viewed, is more like a mathematical exercise in constraint satisfaction, an oft-overlooked but unavoidable aspect of translating abstract ideas into computing machinery. These constraints naturally arise at various stages of syntax validation, type-checking and runtime verification, and help to ensure programs fulfill their intended purpose.

As our work shows, not only is linear algebra over finite fields an expressive language for probabilistic inference, but also an efficient framework for inference on languages themselves. Borrowing analysis techniques from multilinear algebra and tensor completion in the machine learning setting, we develop an equational theory that allows us to translate various decision problems on formal languages into a system of inequalities over finite fields. As a practical consequence, this means we can efficiently encode a number of problems in parsing, code completion and program repair using SAT solvers, which are known to be highly efficient, scalable and flexible to domain-specific constraints. We demonstrate the effectiveness of our approach in a variety of practical scenarios, including code completion and program repair for linear conjunctive languages, and show that our approach is competitive with state-of-the-art methods in terms of both accuracy and efficiency. In contrast with LL and LR-style parsers, our technique can recover partial forests from invalid strings, and handle arbitrary conjunctive languages. In future work, we hope to extend our method to more natural grammars like PCFG, LCFRS and other mildly context-sensitive languages.

Syntax correction tools should be as user-friendly and widely-accessible as autocorrection tools in word processors. From a practical standpoint, we argue it is possible to reduce disruption from manual syntax repair and improve the efficiency of working programmers by driving down the latency needed to synthesize an acceptable repair. In contrast with program synthesizers that require intermediate editor states to be well-formed, our synthesizer does not impose any constraints on the code itself being written and can be used in a live programming environment.

Despite its computational complexity, the design of the tool itself is relatively simple. Tidyparse accepts a linear conjunctive language and a string. If the string is valid, it returns the parse forest, otherwise, it returns a set of repairs, ordered by their perplexity. Tidyparse compiles the grammar and candidate string onto a discrete dynamical system using an extended version of Valiant's algorithm and solves for its fixedpoints using an incremental SAT solver. By allowing the string to contain holes, repairs may contain either concrete tokens or nonterminals, which can be expanded by the user or a neural-guided search procedure to produce a valid program. This approach to parsing has many advantages, enabling us to repair syntax errors, correct typos and recover from errors in real time, as well as being provably sound and complete with respect to the grammatical

specification. It is also compatible with neural program synthesis and repair techniques and shares the same masked language modeling (MLM) target used by Transformer-based LLMs.

We have implemented our approach as an IDE plugin and demonstrated its viability as a practical tool for realtime programming. A considerable amount of effort was devoted to supporting fast error correction functionality. Tidyparse is capable of generating repairs for invalid code in a range of practical languages with very little downstream language integration required. 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.

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] Miltiadis Allamanis, Henry Jackson-Flux, and Marc Brockschmidt. 2021. Self-supervised bug detection and repair. Advances in Neural Information Processing Systems 34 (2021), 27865–27876.

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

[5] Apache Software Foundation. 2022. Apache DataSketches. https://datasketches.apache.org/.

[6] 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.

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

[8] 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

[9] 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).

[10] Richard Beigel and William Gasarch. [n.d.]. A Proof that if $L = L_1 \cap L_2$ where L_1 is CFL and L_2 is Regular then L is Context Free Which Does Not use PDA's. http://www.cs.umd.edu/~gasarch/BLOGPAPERS/cfg.pdf

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

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

[13] 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.

[14] David Chiang, Peter Cholak, and Anand Pillay. 2023. Tighter Bounds on the Expressivity of Transformer Encoders. arXiv preprint arXiv:2301.10743 (2023).

[15] Nadezhda Chirkova and Sergey Troshin. 2021. Empirical study of transformers for source code. In Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. 703–715.

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

[17] 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.

[18] 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

[19] Dawn Drain, Chen Wu, Alexey Svyatkovskiy, and Neel Sundaresan. 2021. Generating bug-fixes using pretrained transformers. In Proceedings of the 5th ACM SIGPLAN International Symposium on Machine Programming. 1–8.

[20] Philippe Duchon, Philippe Flajolet, et al. 2004. Boltzmann samplers for the random generation of combinatorial structures. Combinatorics, Probability and Computing 13, 4-5 (2004), 577–625.

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

[22] Denis Firsov and Tarmo Uustalu. 2015. Certified normalization of context-free grammars. In Proceedings of the 2015 Conference on Certified Programs and Proofs. 167–174.

[23] Seymour Ginsburg and H Gordon Rice. 1962. Two families of languages related to ALGOL. Journal of the ACM (JACM) 9, 3 (1962), 350–371.
[24] Abram Hindle, Earl T Barr, Zhendong Su, Mark Gabel, and Premkumar Devanbu. 2012. On the naturalness of software. In 2012 34th International Conference on Software Engineering (ICSE). IEEE, 837–847. https://doi.org/10.1145/2902362
[25] E. T. Irons. 1963. An Error-Correcting Parse Algorithm. Commun. ACM 6, 11 (nov 1963), 669–673. https://doi.org/10.1145/368310.368385
[26] 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.
[27] 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
[28] William Merrill, Ashish Sabharwal, and Noah A Smith. 2022. Saturated transformers are constant-depth threshold circuits. Transactions of the Association for Computational Linguistics 10 (2022), 843–856.
[29] Matthew Might, David Darais, and Daniel Spiewak. 2011. Parsing with derivatives: a functional pearl. ACM sigplan notices 46, 9 (2011), 189–195.
[30] Alexander Okhotin. 2001. Conjunctive grammars. Journal of Automata, Languages and Combinatorics 6, 4 (2001), 519–535.
[31] Rohit J. Parikh. 1966. On Context-Free Languages. J. ACM 13, 4 (oct 1966), 570–581. https://doi.org/10.1145/321356.321364
[32] Itiroo Sakai. 1961. Syntax in universal translation. In Proceedings of the International Conference on Machine Translation and Applied Language Analysis.
[33] 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.
[34] Klaus U Schulz and Stoyan Mihov. 2002. Fast string correction with Levenshtein automata. International Journal on Document Analysis and Recognition 5 (2002), 67–85.
[35] Michalis K Titsias and Christopher Yau. 2017. The Hamming ball sampler. J. Amer. Statist. Assoc. 112, 520 (2017), 1598–1611.
[36] 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
[37] 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.
[38] Michihiro Yasunaga and Percy Liang. 2021. Break-it-fix-it: Unsupervised learning for program repair. In International Conference on Machine Learning. PMLR, 11941–11952.
[39] Qirun Zhang and Zhendong Su. 2017. Context-sensitive data-dependence analysis via linear conjunctive language reachability. In Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages. 344–358.