

A Survey on String Constraint Solving

ROBERTO AMADINI, University of Bologna

String constraint solving refers to solving combinatorial problems involving constraints over string variables. String solving approaches have become popular over the last years given the massive use of strings in different application domains like formal analysis, automated testing, database query processing, and cybersecurity.

This paper reports a comprehensive survey on string constraint solving by exploring the large number of approaches that have been proposed over the last decades to solve string constraints.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Software and its engineering** → *Constraint and logic languages*; • **Theory of computation** → *Formal languages and automata theory*.

Additional Key Words and Phrases: String Constraint Solving, Constraint Programming, Satisfiability Modulo Theories, Automata Theory, Software Analysis

ACM Reference Format:

Roberto Amadini. 2020. A Survey on String Constraint Solving. 1, 1 (February 2020), 19 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Strings are everywhere across and beyond Computer Science. They are a fundamental datatype in all the modern programming languages, and operations on strings frequently occur in disparate fields such as software analysis, model checking, database applications, web security, bioinformatics and so on [3, 11, 19, 21, 27, 28, 49, 59, 66].

Reasoning over strings requires solving arbitrarily complex *string constraints*, i.e., relations defined on a number of string variables. Typical examples of string constraints are string length, (dis-)equality, concatenation, substring, regular expression matching.

With the term “*string constraint solving*” (in short, string solving or SCS) we refer to the process of modelling, processing, and solving combinatorial problems involving string constraints. We may see SCS as a declarative paradigm which falls at the intersection between constraint solving and combinatorics on words: the user states a problem with string variables and constraints, and a suitable *string solver* seeks a solution for that problem.

Although works on the combinatorics of words were already published in the 1940s [109], the dawn of SCS dates back to the late 1980s in correspondence with the rise of *Constraint Programming* (CP) [113] and *Constraint Logic Programming* (CLP) [72] paradigms. Pioneers in this field were for example Trilogy [141], a language providing strings, integer and real constraints, and CLP(Σ^*) [143], an instance of the CLP scheme representing strings as regular sets. The latter in particular was the first known attempt to use string constraints like regular membership to denote regular sets.

Author’s address: Roberto Amadini, roberto.amadini@unibo.it, University of Bologna, Bologna, Italy.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

Later in the 1990s and 2000s, string solving has sparked some interest (e.g., [37, 63, 69, 80, 82, 100, 108, 110, 124]) without however leaving a mark. It was only from the 2010s that SCS finally took hold in application domains where string processing plays a central role such as test-case generation, software verification, model checking and web security. This increased interest motivated the organisation of the first workshop on string constraints and applications (MOSCA) in 2019 [43].

Arguably, the widespread interest in *cybersecurity* has given new impulse to SCS because strings are often the silent enabler of software vulnerabilities (e.g., SQL injections). Another plausible reason is the remarkable performance improvements that constraint solvers have achieved over the last years. A precise reasoning about strings is especially critical for the analysis of JavaScript language, nowadays the *de-facto* standard for web applications, given the crucial role that strings have in this language [11, 19].

Over the last decade a large number of different SCS approaches has emerged, roughly falling in three main categories:

- (i) *Automata-based approaches*: relying on finite state automata to represent the domain of string variables and to handle string operations.
- (ii) *Word-based approaches*: based on systems of *word equations*. They mainly use *Satisfiability Modulo Theory* (SMT) [24] solvers to tackle string constraints.
- (iii) *Unfolding-based approaches*: they basically expand each string variable x into a number of contiguous elements denoting the characters of x (e.g., x can be mapped into integer variables or bit-vectors).

As we shall see, CP and SMT are the state-of-the-art technologies for solving SCS problems.

The goal of this paper is to provide a comprehensive survey of the various string solving approaches proposed in the literature, ranging from the theoretical foundations to the more practical aspects. After formalising the notion of string constraint solving in Section 2, we provide a detailed review of SCS approaches grouped by category (Sections 3–5). In Section 6 we show the main theoretical results we are aware, while in Section 7 we focus on the practical aspects, by reporting the SCS tools and benchmarks that have been developed. In Section 7 we discuss the related literature before concluding in Section 8.

2 STRING CONSTRAINT SOLVING

In this Section we define the fundamentals of string constraint solving, and in particular we show how SCS are treated from both CP and SMT perspectives. Before that, we recall some preliminary notions about strings and automata theory.

Let us fix a finite *alphabet*, i.e., a set $\Sigma = \{a_1, \dots, a_n\}$ of $n > 1$ symbols also called *characters*. A *string* (or a *word*) w is a finite sequence of $|w| \geq 0$ characters of Σ , and $|w|$ denotes the length of w (in this work we do not consider infinite-length strings). The empty string is denoted with ϵ . The countable set Σ^* of all the strings of Σ is inductively defined as follows: (i) $\epsilon \in \Sigma^*$; (ii) if $a \in \Sigma$ and $w \in \Sigma^*$, then $aw \in \Sigma^*$; (iii) nothing else belongs to Σ^* .

The string *concatenation* of $v, w \in \Sigma^*$ is denoted by $v \cdot w$ (or simply with vw when not ambiguous). We denote with w^n the *iterated concatenation* of w for n times, i.e., $w^0 = \epsilon$ and $w^n = ww^{n-1}$ for $n > 0$. Analogously, we define the concatenation between *sets of strings*: given $V, W \subseteq \Sigma^*$, we denote with $V \cdot W = \{vw \mid v \in V, w \in W\}$ (or simply with VW) their concatenation and with W^n the iterated concatenation, i.e., $W^0 = \{\epsilon\}$ and $W^n = WW^{n-1}$ for $n > 0$.

A set of strings $L \subseteq \Sigma^*$ is called a *formal language*. Formal languages are potentially infinite sets that can be recognised by well-known models of computation called *finite-state automata* (FSA). Roughly, a FSA M is a system

Table 1. Main string constraints.

String constraint	Description
$x = y, x \neq y$	equality, inequality
$x < y, x \leq y, x \geq y, x > y$	lexicographic ordering constraint
$n = x $	string length
$z = x \cdot y, y = x^n$	concatenation, iterated concatenation n times
$y = x^{-1}$	string reversal
$y = x[i..j]$	substring from index i to index j
$n = \text{find}(x, y)$	n is the index of the first occurrence of x in y
$y = \text{replace}(x, q, q')$	y is obtained by replacing the first occurrence of q with q' in x
$y = \text{replaceAll}(x, q, q')$	y is obtained by replacing all the occurrences of q with q' in x
$n = \text{count}(a, x)$	n is the number of occurrences of character a in x
$x \in \mathcal{L}(R)$	membership of x in regular language denoted by R
$x \in \mathcal{L}(G)$	membership of x in context-free grammar language denoted by G

with a finite number of states where a *state transition* occurs according to the input string w . A number $k \geq 0$ of final or accepting states determines if w belongs to the language $\mathcal{L}(M)$ denoted by M or not.

Different variant and extensions of FSA have been proposed, e.g., *deterministic* (DFA), *non-deterministic* (NFA), *push-down automata* (PDA, to recognize context-free languages), *finite-state transducers* (FST, basically FSA with input/output tapes defining relations between sets of strings).

A *string variable* is basically a variable that can only take values in Σ^* or, equivalently, whose domain is a formal language of Σ^* . We can classify string variables into three hierarchical classes:

- *unbounded-length* variables: they can take any value in Σ^*
- *bounded-length* variables: given a fixed integer $\lambda > 0$, they can only take values in $\bigcup_{i=1}^{\lambda} \Sigma^i = \{w \in \Sigma^* \mid |w| \leq \lambda\}$
- *fixed-length* variables: given a fixed $\lambda > 0$, they can only take values in $\Sigma^\lambda = \{w \in \Sigma^* \mid |w| = \lambda\}$

A *string constraint* is a relation over at least a string variable. For example, concatenation is a ternary string constraint. Clearly, instead of writing $(x, y, z) \in \cdot$ we use a more convenient functional notation $z = x \cdot y$. In this paper we will only consider constraints involving strings and (possibly) integers (e.g., string length or iterated concatenation).

Table 1 summarises the main string constraints that one can find in the literature.¹ A good SCS approach should be able to handle most of them. Note that in this work we consider *quantifier-free* constraints only.

From the constraints in Table 1 other string constraints can be derived. For example, the $\text{substring}(x, y)$ constraint imposing that x must be substring of y can be defined as $\text{find}(x, y) > 0$.

Note that Table 1 is arbitrary. For example, the aforementioned $\text{substring}(x, y)$ constraint has the same semantics of $\text{contains}(y, x)$. The $\text{find}(x, y)$ constraint defined in [15] is also referred as $\text{indexOf}(y, x)$. The character indexing is also arbitrary: one SCS approach can index characters starting from 0, while another might prefer a more mathematical 1-based indexing where the first character of string x is $x[1]$ instead of $x[0]$.

In simple terms, the goal of string constraint solving is to determine whether or not a set of string constraints is feasible. As we shall see, this task can be tackled equivalently with both CP and SMT technologies.

¹The count constraint is also referred as *global cardinality count* (GCC) [12].

2.1 SCS from a CP perspective

From a constraint programming point of view, string constraint solving means solving a particular case of *Constraint Satisfaction Problem* (CSP). Formally, a CSP is a triple $P = (\mathcal{X}, \mathcal{D}, C)$ where: $\mathcal{X} = \{x_1, \dots, x_n\}$ are the *variables*; $\mathcal{D} = \{D(x_1), \dots, D(x_n)\}$ are the *domains*, where for $i = 1, \dots, n$ each $D(x_i)$ is a set of values that x_i can take; $C = \{C_1, \dots, C_m\}$ are the *constraints*, i.e., relations over the variables of \mathcal{X} defining the feasible values for the variables.

The goal is to find a *solution* of P , which is basically an assignment $\alpha : \mathcal{X} \rightarrow \bigcup \mathcal{D}$ such that $\alpha(x_i) \in D(x_i)$ for $i = 1, \dots, n$ and $(\alpha(x_{i_1}), \dots, \alpha(x_{i_k})) \in C$ for each constraint $C \in C$ defined over variables $x_{i_1}, \dots, x_{i_k} \in \mathcal{X}$. The CSP notion can be naturally extended to optimization problems: we just add an *objective function* $\varphi : D(x_1) \times \dots \times D(x_n) \rightarrow \mathbb{R}$ that maps each solution to a numerical value to be minimised or maximised.

To find a solution, CP solvers use two main combined techniques: *propagation*, which works on individual constraints trying to prune the domains of the variables involved until a fixpoint is reached, and *branching*, which aims to find a solution via heuristic search (the propagation process is not complete in general).

Fixed an alphabet Σ we call a CSP with strings, or Σ -CSP, a CSP $P = (\mathcal{X}, \mathcal{D}, C)$ having $k > 0$ string variables $\{x_1, \dots, x_k\} \subseteq \mathcal{X}$ such that $D(x_i) \subseteq \Sigma^*$ for $i = 1, \dots, k$, and a number of constraints in C over such variables. To find a solution, a CP solver can try to compile down a Σ -CSP into a CSP with only integer variables [12], or it can define specialised string propagators and branchers [17, 18, 123]. The latter approach has proved to be much more efficient.

For example, consider CSP $P = (\{x, y, n\}, \{\{ab, bc, abcd\}, \Sigma^*, [1, 3]\}, \{n = |x|, x = y^{-1}\})$ where x, y are string variables with associated alphabet $\Sigma = \{a, b, c, d\}$ and n is an integer variable. Propagating $n = |x|$ will exclude string $abcd$ from $D(x)$ because it has length 4, while the domain of n is the interval $[1, 3]$. An optimal propagator for $x = y^{-1}$ would narrow the domain of y from Σ^* to $\{ba, cb\}$. Note that propagation is a compromise between effectiveness (how many values are pruned) and efficiency (the computational cost of pruning), so sometimes it makes sense to settle for efficient but sub-optimal propagators. Then, $n = |x|$ will narrow the domain of n to singleton $\{2\}$ (which actually means assigning the value 2 to n). At this stage, a *fixpoint* is reached, i.e., no more propagation is possible: we have to branch on $\{x, y\}$ to possibly find a solution. Let us suppose that the variable choice heuristics selects variable x and the value choice heuristics assigns to it the value ab ; in this case the propagator of $y = x^{-1}$ is able to conclude that $y = ba$ so a feasible solution for P (not the only one) is $\{x = ab, y = ba, n = 2\}$.

Note that virtually all the CSPs referred in the literature have *finite domains*, i.e., the cardinality of each domain of \mathcal{D} is bounded. Having finite domains guarantees the decidability of CSPs—that are in general NP-complete—by enumeration, but at the same time prevents the use of unbounded-length string variables for Σ -CSPs.

As we shall see in Section 5, all the effective CP approaches for string solving are unfolding-based and do not handle unbounded-length variables. In fact, although CP provides the $\text{regular}(x, M)$ constraint—stating that x must belong to the language denoted by finite state machine M —it is also true that the string variable (or array of integer variables) x must have a fixed [105] or bounded [16] length. The only C(L)P proposals we are aware handling unbounded-length strings (via regular sets) are [143] and [63]. These automata-based approaches are however outdated.

2.2 SCS from a SMT perspective

In a nutshell, Satisfiability Modulo Theories generalises the Boolean satisfiability problem to decide whether a formula in first-order logic is satisfiable with respect to some *background theory* T that fixes the interpretations of predicates and functions [24]. Note that SMT theories can be arbitrarily enriched and combined together.

Over the last decades, several *decision procedures* have been developed to tackle the most disparate theories and sub-theories, including the theory of (non-)linear arithmetic, bit-vectors, floating points, arrays, difference logic, and uninterpreted functions. In particular, well-known SMT solvers like, e.g., CVC4 [88] and Z3 [45] implement the *theory of strings* (often in conjunction with related theories, such as linear arithmetic for length constraints and regular expressions).

For example, the quantifier-free theory T_{SLIA} of strings (or *word equations*) and linear arithmetic deals with integers and unbounded-length strings in Σ^* , where Σ is a given alphabet. Its terms are string/integer variables/constants, concatenation and length. The formulas of T_{SLIA} are (dis-)equalities between strings and linear arithmetic constraints. For example, the formula $\phi \equiv x = ab \cdot z \wedge |x| + |y| \leq 5 \wedge (abcd \cdot x = y \vee |x| > 5)$ where $a, b, c, d \in \Sigma$ and x, y, z are string variables is well-formed for this theory. Unfortunately, the decidability of T_{SLIA} is still unknown [43].

As we shall see in Section 4, over the last years a growing number of modern SMT solvers has integrated the theory of strings. Most of them are based on the DPLL(T) [60] procedure. DPLL(T) is a general framework extending the original DPLL algorithm (tailored for SAT solving) to deal with an arbitrary theory T through the interaction between a SAT solver and a solver specific to T . In a nutshell, DPLL(T) lazily decomposes a SMT problem into a SAT formula, which is handled by a DPLL-based SAT solver which in turn interacts with a theory-specific solver for T , whose job is to check the feasibility of the formulas returned by the SAT solver.

As an example, let us consider the T_{SLIA} theory with the above formula ϕ (which is unsatisfiable). The formula is translated into a Boolean formula ϕ' , handled by a SAT solver which can return “unsatisfiable” or a satisfying assignment (this does not imply that the overall formula is satisfiable). In the latter case, the constraints corresponding to such assignments are distributed to the different theories.

For example, if the assignment $x = ab \cdot z \wedge |x| + |y| \leq 5 \wedge abcd \cdot x = y$ is returned, the constraints $x = ab \cdot z$ and $abcd \cdot x = y$ are delivered to the string solver, while $|x| + |y| \leq 5$ will be solved by an arithmetic solver. Now, the theory solvers can either find that the constraints are T_{SLIA} -satisfiable or return *theory lemmas* to the SAT solver. For example, the string solver might return $\neg(x = ab \cdot z) \vee |x| = |z| + 2$ to the SAT solver, which will add the corresponding clause to its knowledge base. The SAT solver will then produce a new assignment (or return “unsatisfiable”) and the process will be iteratively repeated until either (un-)satisfiability of ϕ is proven or a resource limit is reached (if the theory is not decidable, termination is not guaranteed in general).

3 AUTOMATA-BASED SCS APPROACHES

As aforementioned, the domain of a string variable is a formal language, i.e., a potentially infinite (yet countable) set. A natural way to denote these sets is through (extensions of) finite-state automata. It is therefore unsurprising that the early string solving approaches were based on FSA possibly enriched with other data structures.

We can say that a SCS approach is automata-based if the string variables are principally represented by automata, and the string constraints are mainly mapped into corresponding automata operations.

As aforementioned, CLP(Σ^*) was one of the first attempts to incorporate strings in the CLP framework to strengthen the standard string-handling features such as concatenation and sub-string [143]. This approach was further developed by Golden et al. [63] about 15 years later. Their main contribution was to use FSA to represent regular sets. In [64] Hansen et al. use deterministic FSA (DFA) and *binary decision diagrams* (BDDs) to handle interactive configuration on string variables.

The $\text{regular}(A, M)$ global constraint proposed in [105] treats fixed-size array A of integer variables as a fixed-length string belonging to the regular language denoted by a given (non-)deterministic FSA M . Note that *regular* was introduced to solve finite domains CP problems like rostering and car sequencing, and not targeted to string solving — in fact, it is a useful constraint that has been used in many different CP applications. Its natural extension is the grammar constraint [75, 108], where instead of a FSA we have a context-free grammar. However, grammar constraint never reached the popularity of *regular* in the CP community.

An interesting paper about automata-based approaches is [68], where Hooimeijer et al. study a comprehensive set of algorithms and data structures for automata operations in order to give a fair comparison between different automata-based SCS frameworks [37, 69, 82, 87, 100, 140, 145]. According to their experiments, the best results were achieved when using the BDDs in combination with lazy versions of automata intersection and difference.

MONA [82] is a tool developed in the '90s that acts as a decision procedure for *Monadic Second-Order Logic* (M2L) and as a translator to finite-state automata based on BDDs. FIDO [83] is a domain-specific programming formalism translated first into pure M2L via suitable encodings, and finally into FSA through the MONA tool. Another M2L-based solver is PISA [127], a path- and index-sensitive string solver that is applicable for static analysis.

DRPLE [69] is a SCS approach to solve equations over regular language string variables. The authors provide automata-based decision procedures to tackle the *Regular Matching Assignments* problem and a subclass, the *Concatenation-Intersection* problem.

StrSolve [70] is a decision procedure supporting similar operations to those allowed by DPRLE, but efficiently produces single witnesses rather than atomically generating entire solution sets. Nevertheless, its worst-case performance corresponds to that of DPRLE.

JSA [37] is a string analysis framework that first transforms a Java source into a flow graph (front-end), and then derives FSA from such graph (back-end). In particular, JSA uses well-founded hierarchical directed acyclic graphs of non-deterministic FSA called *multi-level automata* (MLFA).

In [100] Minamide developed a string analyzer for the PHP scripting language to detect cross-site software vulnerabilities and to validate pages they generate dynamically. The analyzer has a library to manipulate formal languages including automata, transducers and context-free grammars.

Rex [140] is a tool based on the Z3 solver [45] for symbolically expressing and analyzing regular expression constraints. It relies on *symbolic finite-state automata* (SFA) where moves are labeled by formulas instead of individual characters. SFAs are then translated into axioms describing the acceptance conditions.

SUSHI [54] is a string solver based on the *Simple Linear String Equation* (SISE) formalism [55] to represent path conditions and attack patterns. To solve SISE constraints, the authors use an automata-based approach. Finite state transducers are used to model the semantics of regular substitution.

Stranger [145] is an automata-based tool for finding and eliminating string-related security vulnerabilities in PHP applications. It uses symbolic forward and backward reachability analyses to compute the possible values that string expressions can take during program execution.

PASS [87] is a string solver using parameterized arrays as the main data structure to model strings, and converts string constraints into quantified expressions that are solved through quantifier elimination. In addition, PASS uses an automaton model to handle regular expressions and reason about string values faster.

SLOG [144] is a string analysis tool based on an NFA manipulation engine with logic circuit representation. Automata manipulations can be performed implicitly using logic circuits while determinization is largely avoided. SLOG also supports symbolic automata and enables the generation of counterexamples.

Sloth [67] is based on the reduction of satisfiability of formulae in the *straight-line* fragment² and in the acyclic fragment to the emptiness problem of *alternating finite-state automata* (AFAs).

Sloth can handle string constraints with concatenation, finite-state transducers, and regular constraints.

OSTRICH [34] is a string solver providing built-in support for concatenation, reverse, functional transducers (FFT), and replaceAll. It can be seen as an extension of Sloth in the sense that the decision algorithm of OSTRICH can reduce the problem to constraints handled by Sloth when it is not possible to avoid non-determinism.

In [149] Zhu et al. proposed a SCS procedure where atomic string constraints are represented by *streaming string transducers* (SSTs) [6]. A straight-line constraint is satisfiable if and only if the domain of the composed streaming string transducer is not empty.

Pros and Cons

Automata enable us to represent infinite sets of strings with finite machines, hence they are a natural and elegant way to represent unbounded-length strings. Also, the theory of automata [142] is well defined and studied.

Unfortunately, the performance of automata-based approaches for string constraint solving has been hampered by two main factors: (i) the possible state explosion due to automata operations (e.g., the intersection of DFA is quadratic in the number of states); (ii) the integration with other domains and theories (integers in particular).

For these reasons, SCS approaches purely based on automata have little success nowadays. However, for some particular classes of SCS problems they still might be the best option.

4 WORD-BASED SCS APPROACHES

As mentioned in Section 2, strings are also called words. What we call string constraints are, in a more algebraic terminology, sometimes called *word equations* [98]. To be more precise, a word equation is a particular string constraint of the form $L = R$ with $L, R \in (\Sigma \cup \mathcal{V})^*$ where Σ is an alphabet and \mathcal{V} is a set of variables.

We can say that a SCS approach is *word-based* if it is based on the theory of word-equations, possibly enriched with other theories (e.g., integers or regular expressions). Word-based approaches rely on algebraic techniques for *natively* solving (quantifier-free) string constraints over the (extended) theory of unbounded strings, without reducing to other data types such as bit vectors or automata. The natural candidates for implementing word-based approaches are SMT solvers, which can incorporate and integrate the theory of strings in their frameworks.

In 2014, Liang et al. [88] argued that: “*Despite their power and success as back-end reasoning engines, general multi-theory SMT solvers so far have provided minimal or no native support for reasoning over strings [...] until very recently the available string solvers were standalone tools that [...] imposed strong restrictions on the expressiveness [...] Traditionally, these solvers were based on reductions to satisfiability problems over other data types*”. However, in the following years the situation has changed, and more and more word-based string solving approaches have emerged.

In [88] the authors integrated a word-based SCS approach into the well-known SMT solver CVC4 [22]. They used a DPLL(T) approach for solving (quantifier-free) constraints natively over the theory of unbounded strings with length and regular language membership. The authors claimed that CVC4 was the first solver able to reason about a language of mixed constraints including strings together with integers, reals, arrays, and algebraic datatypes. This work has been revised and extended in the following years to handle extended string functions frequently occurring in security and verification applications such as contains, indexOf and replace [111, 112].

² The straight-line fragment is based on straight-line conjunctions, which essentially correspond to sequences of program assignments in single static assignment form possibly interleaved with assertions of regular properties [89].

Meanwhile, another well-established SMT solver, namely Z3 [45], started to develop string solving capabilities. Z3-str was introduced in [148] as an extension of Z3 to solve string constraints. Z3-str was the progenitor of a number of different word-based string solvers built on top of Z3. Z3str2 [146] extended Z3-str by including overlapping variables detection and new heuristics. Z3str3 [26] added a technique called theory-aware branching to take into account the structure of theory literals for computing branching activities. Z3strBV [126] is instead a solver for the theory of string equations, string lengths represented as bit-vectors, and bitvector arithmetic aimed at the software verification, testing, and security analysis of C/C++.

Norn is a SMT solver introduced in [4] for an expressive constraint language including word equations, length constraints, and regular membership queries. Norn is based on a decision procedure under the assumption of a set of acyclicity conditions on word equations, where acyclicity is a syntactic condition ensuring that no variable appears more than once in word (dis-)equalities.

Trau [2] is a SMT string solver based on the flattening technique introduced in [1], where *flat automata* are used to capture simple patterns of common constraints. It relies on a Counter-Example Guided Abstraction Refinement (CEGAR) framework [38] where an under- and an over-approximation module interact to increase the string solving precision. In addition, Trau implements string *transduction* by reduction to context-free membership constraints.

S3 is a symbolic string solver [134] motivated by the analysis of web programs working on string inputs. S3 can be viewed as an extension of the aforementioned Z3-str [148] solver to handle regular expressions. Its successor, S3P [135], guides the search towards a “minimal solution” for satisfiable problems and enables conflict clause learning for string theory. The latest version, called S3# [136] implements an algorithm for counting the models of string constraints.

Pros and Cons

State-of-the-art word-based approaches are more general, flexible and often efficient than automata-based approaches. They can be built on the top of well-known SMT solvers and integrated with already defined theories. Furthermore, they can natively handle unbounded-length string.

One the downside, most of these approaches are incomplete and suffer from the performance issues due to the disjunctive reasoning of the underlying DPLL(T) paradigm [60]. In particular, some experimental evaluations [17, 18] showed that they may encounter difficulties when dealing with large string constants.

5 UNFOLDING-BASED SCS APPROACHES

A straightforward way of solving string constraints is to encode them into other well-known types—for which well-established constraint solving techniques already exist—such as Boolean, integers or bit-vectors.

A SCS approach is unfolding-based if each string variable x is unfolded into an homogeneous sequence of $k > 0$ variables of a different type T , and each string constraint is accordingly mapped into a constraint over T .

Note that an unfolding approach inherently needs an *upper bound* λ on the string length. So, these approaches can handle fixed-length or bounded-length string variables, but cannot deal with unbounded-length variables.

A proper choice of λ is crucial. If λ is too small, one cannot capture solutions having not-small-enough string length. On the other hand, too large a value for λ can significantly worsen the SCS performance even for trivial problems.

Another important choice is whether to unfold *eagerly* (i.e., statically, before the actual solving process) or *lazily* (i.e., dynamically, during the solving process).

Hampi [79, 80] was probably the first SMT-based approach encoding string constraints into constraints over bit-vectors, solved by the underlying STP solver [56]. Its first version [80] only allowed one fixed-length string variable.

Its subsequent version [79] added a number of optimisations and, in particular, provided the support to word equations and bounded-length string variables.

Kaluza [117] was the back-end solver used by Kudzu, a symbolic execution framework for the JavaScript code analysis. Similarly to Hampi, Kaluza dealt with string constraints over bounded-length variables by translating them into bit-vector constraints solved with the STP solver [56]. In fact, Kaluza can be seen as an extension of the first version of Hampi [80].

Mapping strings into bit-vectors is suitable for software analysis applications, especially when it comes to precisely handling overflows via *wrapped* integer arithmetic. Plenty of SMT solvers for the quantifier-free bit-vector formulas exist (often relying on *bit blasting*), while the CP support for bit-vectors appears limited [99].

As an alternative to bit-vectors, the simplest approach is to translate string variables into arrays of integer variables encoding the string characters (possibly using a special padding value for the “empty character”). Mapping to integers is probably the most common approach for C(L)P solvers [12].

For example, in [32] the authors describe a lightweight solver relying on CLP and *Constraint Handling Rules* (CHR) paradigm [53] in order to generate large solutions for tractable string constraints in model finding. This approach unfolds string variables by first labelling their lengths and domains, and then their characters.

As previously mentioned, CP solvers can use fixed-length or bounded-length arrays of integer variables to deal with regular, grammar and other string constraints [12, 65, 76, 97, 105, 108] without a native support for string variables. However, as shown in [12, 17, 18, 123], having dedicated propagators for string variables can make a difference.

In [120] Scott et al. presented a prototypical bounded-length approach based on the *affix* domain to natively handle string variables. This domain allows one to reason about the content of string suffixes even when the length is unknown by using a padding symbol at the end of the string.

The [120] approach has been subsequently improved [119, 122, 123] with a new structured variable type for strings called *Open-Sequence Representation*, for which suitable propagators are defined. This approach has been implemented in the Gecode solver [61] and in [12] referred as GECODE+S.

To mitigate the dependency on λ and enable lazy unfolding, a fairly recent CP approach based on *dashed strings* has been introduced [17, 18]. Dashed strings are concatenations of distinct set of strings (called blocks) used to represent in a compact way the domain of string variables with potentially very big length. Dashed string propagators have been defined for a number of constraints [15–17] and implemented in the G-Strings solver [10].

Pros and cons

The unfolding-based approaches allows one to take advantage of already defined theories and propagators without explicitly implementing support for strings. Experimental results show that unfolding approaches, and in particular the CP-based dashed string approach, can be quite effective — especially for SCS problems involving long strings.

However, unfolding approaches also have limitations. The most obvious one is the impossibility of handling unbounded-length strings. This can be negligible, provided that a good value of λ is chosen. Unfortunately, deciding a good value of λ is not always trivial. CP solvers can be more efficient than SMT solvers for satisfiable SCS problems, but they may fail on unsatisfiable problems with large domains, and on problems with a lot of logical disjunctions.

6 THEORETICAL ASPECTS

In this section we briefly explore the theoretical side of string constraint solving, focusing in particular on the results that have been applied to the SMT and CP fields. For more insights about the theory of automata and combinatorics on words we refer the reader to [25, 35, 42, 62, 71, 92–94, 114, 142].

Algebraically speaking, the structure $(\Sigma^*, \cdot, \epsilon)$ is a *free monoid* on Σ . The *free semigroup* on Σ is instead the semigroup (Σ^+, \cdot) where $\Sigma^+ = \Sigma^* - \{\epsilon\}$. There is therefore a number of theoretical results for free semigroups that can be applied for string solving [48, 98, 138].

Over the years the complexity and the decidability of combinatorial problems over finite-length strings has been deeply studied [31, 48, 52, 77, 90, 98, 107, 109, 138]. Much progress has been made, but many questions remain open – especially when the language is enriched with new predicates [57]. For example, it is still unknown whether the theory of word equations and arithmetic over length functions is decidable [?]. In [102], Néraud provided a list of open problems and conjectures for combinatorics on words.

As also noted in [5], DPLL(T)-based string solvers like Z3str2 [147], Z3str3 [26], CVC4 [88], S3 [134–136], Norn [4], Trau [2], Sloth [67], and OSTRICH [34] handle a variety of string constraints but they are *incomplete* for the full combination of those constraints. So, they often decide to deal with a particular fragment of the individual string constraints.

In [57] Ganesh et al. proved several (un-)decidability results for word equations with length constraints and membership in regular sets. They proved in particular three main theorems: (i) the undecidability of the validity problem for the set of sentences written as a $\forall\exists$ quantifier alternation applied to positive word equations; (ii) the decidability of quantifier-free formulas over word equations in solved form and length constraints; (iii) the decidability of quantifier-free formulas over word equations in regular solved form, length constraints, and the membership predicate over regular expressions.

The work in [44] extends [57] by considering the (un-)decidability of many fragments of the first order theory of word equations and their extensions. The authors show that when extended with several natural predicates on words, the existential fragment becomes undecidable. Moreover, they prove that deciding whether solutions exist for a restricted class of equations, augmented with many of the predicates leading to undecidability in the general case, is possible in non-deterministic polynomial time. The undecidability of the theory of strings with linear arithmetic over length and string-number conversion is proven in [?].

A. W. Lin et al. [89] studied the decidability of string logics with concatenations and finite-state transducers as atomic operations. They show that the straight-line fragment of the logic is decidable (complexity ranges from PSPACE to EXPSpace). This fragment can express constraints required for analysing mutated XSS (mXSS) in web applications and remains decidable in the presence of length, letter-counting, regular, indexOf, and disequality constraints.

In [33] Chen et al. provide a systematic study of straight-line string constraints with replace and regular membership as basic operations. They show that a large class of such constraints (i.e., when only a constant string or a regular expression is permitted in the pattern) is decidable. They also show undecidability results when variables or length constraints are permitted in the pattern parameter of the replace function.

In [34] Chen et al. define a first-order language based on the .NET string library functions and prove some decidability properties for (fragments of) that language. In particular, based on the work of [31], the authors prove that undecidability of the path feasibility problem for a fragment called library language.

In [5] Abdulla et al. address string constraints combining transducers, word equations, and length constraints. Since this problem is undecidable in general, they propose a new decidable fragment of string constraints, called *weakly chaining*, that they prove to be decidable.

Recompression [74] is a technique based on local modification of variables and iterative replacement of pairs of letters that can be seen as a bottom-up compression of the solution of a given word equation. Recompression can be used for the proof of satisfiability of word equations [74] or for grammar-based compression [73].

Switching to the CP side, we cannot say anything deep about decidability because this paradigm essentially works on finite domains, and CSPs over finite domains are trivially decidable by enumerating the domain values. The theoretical results in CP mainly refer to the *consistency* levels [113] that a given constraint propagator can enforce. For example, the regular propagator introduced in [105] enforces *Generalized Arc Consistency* (GAC).

In [106] Pesant et al. performed a theoretical study demonstrating how the grammar constraint can be linearized efficiently. In particular, they proposed a lifted polytope having only integer extreme points.

In [76] Kadioglu et al. theoretically investigate filtering problems that arise from grammar constraints. They authors address question like: can we efficiently filter context-free grammar constraints? How can we achieve arc-consistency for conjunctions of regular grammar constraints? What languages are suited for filtering based on regular and context-free grammar constraints? Are there languages that are suited for context-free, but not for regular grammar filtering?

In [121] Scott et al. extended the open domain consistency notion of [97] with new domain-specific definitions (PS-consistency, PSL-consistency, PSU-consistency, and PSLU-consistency). These notions are however incomparable with the consistency level reached [122].

Given their non-lattice nature, dashed strings do not come with propagators that formally ensure a given level of consistency. The main theoretical results for dashed strings concern the sweep-based equation algorithm introduced in [18], on which most of the dashed string propagators rely, which has been proven to be sound (while its completeness is still an open problem).

7 PRACTICAL ASPECTS

In this Section we focus on the practical aspects of string solving, that is, on the SCS tools and benchmarks that have been actually developed and tested.

Among the automata-based approaches seen in Section 3, the tools that are currently available online are MONA [129], JSA [128], the PHP string analyzer [101] described in [100], Rex [139], and Sloth [51]. However, only the latter two are actively maintained. PISA [127] is a commercial security product not publicly available.

CVC4 and Z3 are two of the most famous SMT solvers. Their interface for string solving is maintained and well documented [130, 131]. Note that Z3 provides two alternatives for string solving: the theory of strings (via Z3str3 solver [26]) and the theory of sequences (Z3seq). Z3str2 and Z3-str are no longer maintained.

Norn is publicly available at [125], but its last update dates back to 2015. At [137] one can find the binaries of all the three versions of S3 string solver (S3, S3P, and S3#). However, the sources are available only for S3 and no longer updated since 2014.

Trau comes in two versions. The first one is available at [46], while the newest implementation (called Z3-Trau) can be found at [47]. Both these versions are built on top of the Z3 solver.

Table 2. StringFuzz benchmarks composition

Class	Description	Quantity
<i>Concats-{Small,Big}</i>	Right-heavy, deep tree of concats	120
<i>Concats-Balanced</i>	Balanced, deep tree of concats	100
<i>Concats-Extracts-{Small,Big}</i>	Single concat tree, with character extractions	120
<i>Lengths-{Long,Short}</i>	Single, large length constraint on a variable	200
<i>Lengths-Concats</i>	Tree of fixed-length concats of variables	100
<i>Overlaps-{Small,Big}</i>	Formula of the form $ax = xb$	80
<i>Regex-{Small,Big}</i>	Complex regex membership test	120
<i>Many-Regexes</i>	Multiple random regex membership tests	40
<i>Regex-Deep</i>	Regex membership test with many nested operators	45
<i>Regex-Pair</i>	Test for membership in one regex, but not another	40
<i>Regex-Lengths</i>	Regex membership test, and a length constraint	40
<i>Different-Prefix</i>	Equality of two deep concats with different prefixes	60

Concerning unfolding-based approaches, Kaluza solver is still available at [116] but no longer maintained. Analogously, GECODE+S is available on-line [118] but no longer developed. Conversely, the G-Strings solver [9] implementing the dashed string approach is actively maintained. Both GECODE+S and G-Strings are extensions of Gecode [61], a well-established CP solver over finite domains.

ACO-Solver [132] is a tool using a hybrid constraint solving procedure based on the *ant colony optimization* meta-heuristic, which is executed as a fall-back mechanism together with an automata-based solver, when an underlying solver encounters an unsupported string operation (e.g., sanitization functions or regular expression matching). JOACO-CS is an extension of ACO-Solver publicly available at [50].

All the above approaches, and many others, have been tested on several string benchmarks — especially coming from the software verification world. Although some of them are publicly available, a major issue for string solving is that there are no standard benchmarks for SCS applications. This makes difficult a rigorous comparison and the integration between string solvers.

The first step towards the creation of standard benchmarks is the implementation of a *common language* to model and solve SCS problems. Unfortunately, at present, strings are not a standard feature for neither SMT nor CP modelling languages. However, there is movement in this direction.

For example, given the increasing interest in string solving, the STM-LIB initiative [23] is developing an official theory of strings and related logics for SMT solvers.³ On the CP side, the de-facto modelling language for constraint solvers is called *MiniZinc* [103]. Strings are not (yet) part of the official MiniZinc release, but in [12] Amadini et al. introduced an extension to enable string solving. This also includes a library for compiling string variables and constraints into bounded-length arrays or integer variables and constraints.

Unfortunately, at the moment the interaction between the SMT and CP communities does not appear very strong. A possible way to bring these two communities closer together is to define compilers to translate SMT-LIB into MiniZinc and vice versa. For example, Bofill et al. [30] defined a compiler from FlatZinc—the low-level language derived from MiniZinc—to SMT-LIB. More recently, G. Gange [58] implemented a prototypical compiler from SMT-LIB to MiniZinc able to process string variables and constraints.

³<http://smtlib.cs.uiowa.edu/theories-UnicodeStrings.shtml>

A nice step forward for SCS benchmarks is represented by *StringFuzz* [29], a modular SMT-LIB problem instance transformer and generator for string solvers. It is a useful tool for string solver developers and testers, and can help expose bugs and performance issues.

A repository of several SMT-LIB 2.0/2.5 problem instances generated and transformed with StringFuzz is available online⁴. Table 2 summarizes the nature of these instances, grouped into twelve different classes. StringFuzz also reports the performance of Z3str3, CVC4, Z3seq, and Norn on such instances. Only these SMT solvers were used because other SMT approaches are either unstable or cannot properly process the proposed SMT-LIB syntax. Note that the SMT-LIB to MiniZinc compiler of [58] can successfully translate all the generated StringFuzz instances, so also the performance of CP solvers can be evaluated.

In this section we cover some of the related literature that has not been previously mentioned.

DReX [7] is a declarative language that can express all the regular string-to-string transformations based on function combinators [8]. The focus is on the complexity of evaluating the output of a DReX program on a given input string. In particular, the main contribution of [7] is the identification of a consistency restriction on the use of combinators in DReX programs, and a single-pass evaluation algorithm for consistent programs.

Luu et al. developed SMC [95], a *model counter* for determining the number of solutions of combinatorial problems involving string constraints. Their work relies on *generating functions*, a mathematical tool for reasoning about infinite series that also provides a mechanism to handle the cardinality of string sets. SMC is expressive enough to model constraints arising in real-world JavaScript applications and UNIX C utilities.

An Automata-Based model Counter for string constraints (ABC) is implemented in [20]. ABC uses an automata-based constraint representation for reducing model counting to path counting. In [20] the ABC model counter is extended with relational and numeric constraints.

As previously mentioned, the main application field for string solving is the area of software verification and testing. For example, Kausler et al. [78] performed an evaluation of string constraint solvers in the context of symbolic execution [81]. What they find is that, as one can expect, one solver might be more appropriate than another depending on the input program. This is also pointed out in [11], where Amadini et al. presented a multi-solver tool for the dynamic symbolic execution of JavaScript.

In [115], the PHPRepair tool is used for automatically repairing HTML generation errors in PHP via string constraint solving. The property that all tests of a suite should produce their expected output is modelled with string constraints encoded into the language of Kodkod [133], a SAT-based constraint solver.

Abstracting a set of strings with a finite formalism is not merely a string solving affair. For example, the well-known *Abstract Interpretation* [41] framework may require the sound approximation of sets of strings (i.e., all the possible “concrete” values that a string variable of the input program can take) with an abstract counterpart.

Several abstract domains have been proposed to approximate set of strings. These domains vary according to the properties that one needs to capture (e.g., the string length, the prefix or suffix, the characters occurring in a string). Madsen et al. [96] proposed and evaluated a suite of twelve string domains for the static analysis of dynamic field access. Additional string domains are also discussed in [39, 40].

Choi et al. [36] used restricted regular expressions as an abstract domain for strings in the context of Java analysis. Park et al. [104] use a stricter variant of this idea, with a more clearly defined string abstract domain.

⁴<http://stringfuzz.dmitryblotsky.com/benchmarks/>

Amadini et al. [19] provided an evaluation on the combination via *direct product* of different string abstract domains in order to improve the precision of JavaScript static analysis. In [13] this combination is achieved via *reduced product* by using the set of regular languages as a *reference domain* for the other string domains.

Finally, we mention that string solving techniques might be useful in the context of *Bioinformatics*, where a number of CP techniques have been already applied (see, e.g., the works by Barahona et al. [21, 85, 86]).

8 CONCLUSIONS

In this work we provided a comprehensive survey on the various aspects of string constraint solving (SCS), an emerging important field orthogonal to combinatorics on words and constraint solving. In particular, we focused on the three main categories of the SCS approaches we are aware (automata-based, word-based and unfolding-based) from the early proposals to the state-of-the-art approaches.

Among the future directions for string constraint solving we mention the four challenges reported in [14], namely:

- *Extending* SCS capabilities to properly handle complex string operations, frequently occurring in web programming [91], such as *back-references*, *lookaheads/lookbehinds* or *greedy matching*.
- *Improving* the efficiency of SCS solvers with new algorithms and search heuristics. At present, SMT solvers tend to fail with long-length strings, while CP solvers may struggle to prove unsatisfiability.
- *Combining* SCS solvers with a *portfolio approach* [84] in order to exploit their different nature and uneven performance across different problem instances.
- *Using* SCS solvers and related tools in different fields. The best candidates are probably software verification and testing, model checking and cybersecurity.

Finally, we hope that the research in SCS will encourage a closer and more fruitful collaboration between the CP and the SAT/SMT communities.

REFERENCES

- [1] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Bui Phi Diep, Lukás Holík, Ahmed Rezine, and Philipp Rümmer. 2017. Flatten and Conquer: A Framework for Efficient Analysis of String Constraints. In *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2017*. 602–617.
- [2] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Bui Phi Diep, Lukás Holík, Ahmed Rezine, and Philipp Rümmer. 2018. Trau: SMT solver for string constraints. In *2018 Formal Methods in Computer Aided Design, FMCAD 2018, Austin, TX, USA, October 30 - November 2, 2018*, Nikolaj Bjørner and Arie Gurfinkel (Eds.). IEEE, 1–5.
- [3] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Lukás Holík, Ahmed Rezine, Philipp Rümmer, and Jari Stenman. 2014. String Constraints for Verification. In *Computer Aided Verification - 26th International Conference, CAV 2014, Held as Part of the Vienna Summer of Logic, VSL 2014, Vienna, Austria, July 18-22, 2014. Proceedings (Lecture Notes in Computer Science)*, Armin Biere and Roderick Bloem (Eds.), Vol. 8559. Springer, 150–166.
- [4] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Lukás Holík, Ahmed Rezine, Philipp Rümmer, and Jari Stenman. 2015. Norn: An SMT Solver for String Constraints. In *CAV (LNCS)*, Vol. 9206. Springer, 462–469.
- [5] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Bui Phi Diep, Lukás Holík, and Petr Janku. 2019. Chain-Free String Constraints. In *Automated Technology for Verification and Analysis - 17th International Symposium, ATVA 2019, Taipei, Taiwan, October 28-31, 2019, Proceedings (Lecture Notes in Computer Science)*, Yu-Fang Chen, Chih-Hong Cheng, and Javier Esparza (Eds.), Vol. 11781. Springer, 277–293.
- [6] Rajeev Alur. 2011. Streaming String Transducers. In *Logic, Language, Information and Computation*, Lev D. Beklemishev and Ruy de Queiroz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–1.
- [7] Rajeev Alur, Loris D’Antoni, and Mukund Raghothaman. 2015. DReX: A Declarative Language for Efficiently Evaluating Regular String Transformations. In *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2015, Mumbai, India, January 15-17, 2015*, Sriram K. Rajamani and David Walker (Eds.). ACM, 125–137.

- [8] Rajeev Alur, Adam Freilich, and Mukund Raghothaman. 2014. Regular combinators for string transformations. In *Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS), CSL-LICS '14, Vienna, Austria, July 14 - 18, 2014*, Thomas A. Henzinger and Dale Miller (Eds.). ACM, 9:1–9:10.
- [9] Roberto Amadini. [n.d.]. G-Strings - Gecode with (dashed) string variables. Available at <https://github.com/ramadini/gecode>.
- [10] Roberto Amadini. 2020. G-Strings: Gecode with (dashed) string variables. Available at <https://github.com/ramadini/gecode>.
- [11] Roberto Amadini, Mak Andron, Graeme Gange, Peter Schachte, Harald Søndergaard, and Peter J. Stuckey. 2019. Constraint Programming for Dynamic Symbolic Execution of JavaScript. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research - 16th International Conference, CPAIOR 2019, Thessaloniki, Greece, June 4-7, 2019, Proceedings (Lecture Notes in Computer Science)*, Louis-Martin Rousseau and Kostas Stergiou (Eds.), Vol. 11494. Springer.
- [12] Roberto Amadini, Pierre Flener, Justin Pearson, Joseph D. Scott, Peter J. Stuckey, and Guido Tack. 2017. MiniZinc with strings. In *LOPSTR 2016: Revised Selected Papers (LNCS)*, Manuel Hermenegildo and Pedro López-García (Eds.), Vol. 10184. Springer, 59–75.
- [13] Roberto Amadini, Graeme Gange, François Gauthier, Alexander Jordan, Peter Schachte, Harald Søndergaard, Peter J. Stuckey, and Chenyi Zhang. 2018. Reference Abstract Domains and Applications to String Analysis. *Fundam. Inform.* 158, 4 (2018), 297–326.
- [14] Roberto Amadini, Graeme Gange, Peter Schachte, Harald Søndergaard, and Peter J. Stuckey. 2020. String constraint solving: past, present and future. To appear in the proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020).
- [15] Roberto Amadini, Graeme Gange, and Peter J. Stuckey. 2018. Propagating *Lex*, *Find* and *Replace* with Dashed Strings. In *Proc. 15th Int. Conf. Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (LNCS)*, W.-J. van Hoeve (Ed.), Vol. 10848. Springer.
- [16] Roberto Amadini, Graeme Gange, and Peter J. Stuckey. 2018. Propagating Regular Membership with Dashed Strings. In *Proc. 24th Conf. Principles and Practice of Constraint Programming (LNCS)*, J. Hooker (Ed.), Vol. 11008. Springer, 13–29.
- [17] Roberto Amadini, Graeme Gange, and Peter J. Stuckey. 2018. Sweep-Based Propagation for String Constraint Solving. In *Proc. 32nd AAAI Conf. Artificial Intelligence*. AAAI Press, 6557–6564.
- [18] Roberto Amadini, Graeme Gange, Peter J. Stuckey, and Guido Tack. 2017. A Novel Approach to String Constraint Solving. In *Proc. 23rd Int. Conf. Principles and Practice of Constraint Programming (LNCS)*, J. C. Beck (Ed.), Vol. 10416. Springer, 3–20.
- [19] Roberto Amadini, Alexander Jordan, Graeme Gange, François Gauthier, Peter Schachte, Harald Søndergaard, Peter J. Stuckey, and Chenyi Zhang. 2017. Combining String Abstract Domains for JavaScript Analysis: An Evaluation. In *Tools and Algorithms for the Construction and Analysis of Systems - 23rd International Conference, TACAS 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22-29, 2017, Proceedings, Part I*. 41–57.
- [20] Abdulkali Aydin, Lucas Bang, and Tevfik Bultan. 2015. Automata-Based Model Counting for String Constraints. In *Computer Aided Verification - 27th International Conference, CAV 2015, San Francisco, CA, USA, July 18-24, 2015, Proceedings, Part I (Lecture Notes in Computer Science)*, Daniel Kroening and Corina S. Pasareanu (Eds.), Vol. 9206. Springer, 255–272.
- [21] Pedro Barahona and Ludwig Krippahl. 2008. Constraint Programming in Structural Bioinformatics. *Constraints* 13, 1-2 (2008), 3–20.
- [22] Clark Barrett, Christopher L. Conway, Morgan Deters, Liana Hadarean, Dejan Jovanović, Tim King, Andrew Reynolds, and Cesare Tinelli. 2011. CVC4. In *Computer Aided Verification*, Ganesh Gopalakrishnan and Shaz Qadeer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 171–177.
- [23] Clark Barrett, Pascal Fontaine, and Cesare Tinelli. 2016. The Satisfiability Modulo Theories Library (SMT-LIB). Available at <https://www.SMT-LIB.org>.
- [24] Clark W. Barrett and Cesare Tinelli. 2018. Satisfiability Modulo Theories. In *Handbook of Model Checking*. 305–343.
- [25] Jean Berstel and Dominique Perrin. 2007. The origins of combinatorics on words. *European Journal of Combinatorics* 28, 3 (2007), 996–1022.
- [26] Murphy Berzish, Vijay Ganesh, and Yunhui Zheng. 2017. Z3str3: A String Solver with Theory-Aware Heuristics. In *Proc. 17th Conf. Formal Methods in Computer-Aided Design*, D. Stewart and G. Weissenbacher (Eds.). FMCAD Inc, 55–59.
- [27] Prithvi Bisht, Timothy L. Hinrichs, Nazari Skrupsky, and V. N. Venkatakrishnan. 2011. WAPTEC: Whitebox analysis of web applications for parameter tampering exploit construction. In *Proceedings of ACM Conference on Computer and Communications Security*. ACM, 575–586.
- [28] Nikolaj Bjørner, Nikolai Tillmann, and Andrei Voronkov. 2009. Path Feasibility Analysis for String-Manipulating Programs. In *Proceedings of the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (LNCS)*, Vol. 5505. Springer, 307–321.
- [29] Dmitry Blotsky, Federico Mora, Murphy Berzish, Yunhui Zheng, Ifaz Kabir, and Vijay Ganesh. 2018. StringFuzz: A Fuzzer for String Solvers. In *Computer Aided Verification*, Hana Chockler and Georg Weissenbacher (Eds.). Springer International Publishing, Cham, 45–51.
- [30] Miquel Bofill, Josep Suy, and Mateu Villaret. 2010. A system for solving constraint satisfaction problems with SMT. In *SAT (LNCS)*, Vol. 6175. Springer, 300–305.
- [31] J. Richard Büchi and Steven Senger. 1990. *Definability in the Existential Theory of Concatenation and Undecidable Extensions of this Theory*. Springer New York, New York, NY, 671–683.
- [32] Fabian Büttner and Jordi Cabot. 2015. Lightweight string reasoning in model finding. *Software and Systems Modeling* 14, 1 (2015), 413–427.
- [33] Taolue Chen, Yan Chen, Matthew Hague, Anthony W. Lin, and Zhilin Wu. 2018. What is decidable about string constraints with the ReplaceAll function. *PACMPL* 2, POPL (2018), 3:1–3:29.
- [34] Taolue Chen, Matthew Hague, Anthony W. Lin, Philipp Rümmer, and Zhilin Wu. 2019. Decision procedures for path feasibility of string-manipulating programs with complex operations. *PACMPL* 3, POPL (2019), 49:1–49:30.
- [35] Christian Choffrut and Juhani Karhumäki. 1997. Combinatorics of words. In *Handbook of formal languages*. Springer, 329–438.

- [36] Tae-Hyoung Choi, Oukseh Lee, Hyunha Kim, and Kyung-Goo Doh. 2006. A Practical String Analyzer by the Widening Approach. In *Programming Languages and Systems, 4th Asian Symposium, APLAS 2006, Sydney, Australia, November 8-10, 2006, Proceedings (Lecture Notes in Computer Science)*, Naoki Kobayashi (Ed.), Vol. 4279. Springer, 374–388.
- [37] Aske Simon Christensen, Anders Møller, and Michael I. Schwartzbach. 2003. Precise Analysis of String Expressions. In *SAS (LNCS)*, Vol. 2694. Springer, 1–18.
- [38] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. 2000. Counterexample-Guided Abstraction Refinement. In *Computer Aided Verification*, E. Allen Emerson and Aravinda Prasad Sistla (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 154–169.
- [39] Agostino Cortesi and Martina Oliaro. 2018. M-String Segmentation: A Refined Abstract Domain for String Analysis in C Programs. In *2018 International Symposium on Theoretical Aspects of Software Engineering, TASE 2018, Guangzhou, China, August 29-31, 2018*, Jun Pang, Chenyi Zhang, Jifeng He, and Jian Weng (Eds.). IEEE Computer Society, 1–8.
- [40] Giulia Costantini, Pietro Ferrara, and Agostino Cortesi. 2015. A suite of abstract domains for static analysis of string values. *Software: Practice and Experience* 45, 2 (2015), 245–287.
- [41] Patrick Cousot and Radhia Cousot. 1977. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. In *Proceedings of the Fourth ACM Symposium on Principles of Programming Languages*. ACM, 238–252.
- [42] M. Dal Cin. 1980. *The Algebraic Theory of Automata*. Vieweg+Teubner Verlag, Wiesbaden, 348–361.
- [43] Loris D’Antoni, Anthony W. Lin, and Philipp R  ijmmer. 2019. Meeting on String Constraints and Applications. <https://mosca19.github.io/>.
- [44] Joel D. Day, Vijay Ganesh, Paul He, Florin Manea, and Dirk Nowotka. 2018. The Satisfiability of Word Equations: Decidable and Undecidable Theories. In *Reachability Problems - 12th International Conference, RP 2018, Marseille, France, September 24-26, 2018, Proceedings (Lecture Notes in Computer Science)*, Igor Potapov and Pierre-Alain Reynier (Eds.), Vol. 11123. Springer, 15–29.
- [45] Leonardo Mendon  a de Moura and Nikolaj Bj  rner. 2008. Z3: An Efficient SMT Solver. In *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2008, Budapest, Hungary, March 29-April 6, 2008. Proceedings*. 337–340.
- [46] Bui Phi Diep. [n.d.]. Trau. Available at <https://github.com/diepbp/Trau>.
- [47] Bui Phi Diep. [n.d.]. Z3-Trau. Available at https://github.com/guluchen/z3/tree/new_trau.
- [48] Valery G. Durnev. 1995. Undecidability of the positive $\forall\exists$ -theory of a free semigroup. *Siberian Mathematical Journal* 36 (1995), 917–929.
- [49] Michael Emmi, Rupak Majumdar, and Koushik Sen. 2007. Dynamic Test Input Generation for Database Applications. In *Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 151–162.
- [50] Julian Thom   et al. [n.d.]. Joaco. Available at <https://sites.google.com/site/joacosite/>.
- [51] Luk  s Hol  k et al. [n.d.]. Sloth - An SMT Solver for String Constraints. Available at <https://github.com/uuverifiers/sloth>.
- [52] Ann S. Ferebee. 1972. Review: Yu. V. Matiyasevich, A. O. Slisenko, The Connection between Hilbert’s Tenth Problem and Systems of Equations between Words and Lengths. *J. Symbolic Logic* 37, 3 (09 1972), 604. <https://projecteuclid.org:443/euclid.jsl/1183738340>
- [53] Thom W. Fr  hwirth. 1998. Theory and Practice of Constraint Handling Rules. *J. Log. Program.* 37, 1-3 (1998), 95–138.
- [54] Xiang Fu and Chung-Chih Li. 2010. A String Constraint Solver for Detecting Web Application Vulnerability. In *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE’2010), Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010*. Knowledge Systems Institute Graduate School, 535–542.
- [55] Xiang Fu, Michael C. Powell, Michael Bantegui, and Chung-Chih Li. 2013. Simple linear string constraints. *Formal Asp. Comput.* 25, 6 (2013), 847–891. <https://doi.org/10.1007/s00165-011-0214-3>
- [56] Vijay Ganesh and David L. Dill. 2007. A Decision Procedure for Bit-Vectors and Arrays. In *Computer Aided Verification, 19th International Conference, CAV 2007, Berlin, Germany, July 3-7, 2007, Proceedings (Lecture Notes in Computer Science)*, Werner Damm and Holger Hermanns (Eds.), Vol. 4590. Springer, 519–531.
- [57] Vijay Ganesh, Mia Minnes, Armando Solar-Lezama, and Martin C. Rinard. 2012. Word Equations with Length Constraints: What’s Decidable?. In *Hardware and Software: Verification and Testing - 8th International Haifa Verification Conference, HVC 2012, Haifa, Israel, November 6-8, 2012. Revised Selected Papers (Lecture Notes in Computer Science)*, Armin Biere, Amir Nahir, and Tanja E. J. Vos (Eds.), Vol. 7857. Springer, 209–226.
- [58] Graeme Gange. [n.d.]. smt2mzn-str. Available at <https://bitbucket.org/gkgange/smt2mzn-str>.
- [59] Graeme Gange, Jorge A. Navas, Peter J. Stuckey, Harald S  ndergaard, and Peter Schachte. 2013. Unbounded Model-Checking with Interpolation for Regular Language Constraints. In *Proceedings of the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (LNCS)*, Vol. 7795. Springer, 277–291.
- [60] H. Ganzinger, G. Hagen, R. Nieuwenhuis, A. Oliveras, and C. Tinelli. 2004. DPLL(T): Fast Decision Procedures. In *Computer Aided Verification: Proc. 16th Int. Conf. (LNCS)*, R. Alur and D. A. Peled (Eds.), Vol. 3114. Springer, 175–188.
- [61] Gecode Team. 2016. Gecode: Generic Constraint Development Environment. Available at <http://www.gecode.org>.
- [62] Victor M. Glushkov. 1961. The Abstract Theory of Automata. *Russian Mathematical Surveys* 16, 5 (1961), 1–53.
- [63] Keith Golden and Wanlin Pang. 2003. Constraint Reasoning over Strings. In *Proceedings of the Ninth International Conference on Principles and Practice of Constraint Programming*. 377–391.
- [64] Esben Rune Hansen and Henrik Reif Andersen. 2007. Interactive Configuration with Regular String Constraints. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*. AAAI Press, 217–223.

- [65] Jun He, Pierre Flener, Justin Pearson, and Wei Ming Zhang. 2013. Solving String Constraints: The Case for Constraint Programming. In *Principles and Practice of Constraint Programming - 19th International Conference, CP 2013, Uppsala, Sweden, September 16-20, 2013. Proceedings*, Christian Schulte (Ed.), Vol. 8124. Springer, 381–397.
- [66] Hossein Hojjat, Philipp Rümmer, and Ali Shamakhi. 2019. On Strings in Software Model Checking. In *Programming Languages and Systems - 17th Asian Symposium, APLAS 2019, Nusa Dua, Bali, Indonesia, December 1-4, 2019, Proceedings (Lecture Notes in Computer Science)*, Anthony Widjaja Lin (Ed.), Vol. 11893. Springer, 19–30.
- [67] Lukás Holík, Petr Janku, Anthony W. Lin, Philipp Rümmer, and Tomáš Vojnar. 2018. String constraints with concatenation and transducers solved efficiently. *PACMPL* 2, POPL (2018), 4:1–4:32.
- [68] Pieter Hooimeijer and Margus Veanes. 2011. An Evaluation of Automata Algorithms for String Analysis. In *VMCAI (LNCS)*, Vol. 6538. Springer, 248–262.
- [69] Pieter Hooimeijer and Westley Weimer. 2009. A decision procedure for subset constraints over regular languages. In *Proceedings of the 2009 ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2009, Dublin, Ireland, June 15-21, 2009*, Michael Hind and Amer Diwan (Eds.). ACM, 188–198.
- [70] Pieter Hooimeijer and Westley Weimer. 2012. StrSolve: Solving string constraints lazily. *Automated Software Engineering* 19, 4 (2012), 531–559.
- [71] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2007. *Introduction to automata theory, languages, and computation, 3rd Edition*. Addison-Wesley.
- [72] Joxan Jaffar and Michael J. Maher. 1994. Constraint Logic Programming: A Survey. *J. Log. Program.* 19/20 (1994), 503–581. [https://doi.org/10.1016/0743-1066\(94\)90033-7](https://doi.org/10.1016/0743-1066(94)90033-7)
- [73] Artur Jez. 2015. Approximation of grammar-based compression via recompression. *Theor. Comput. Sci.* 592 (2015), 115–134.
- [74] Artur Jez. 2016. Recompression: A Simple and Powerful Technique for Word Equations. *J. ACM* 63, 1 (2016), 4:1–4:51. <https://doi.org/10.1145/2743014>
- [75] Serdar Kadioglu and Meinolf Sellmann. 2010. Grammar constraints. *Constraints* 15, 1 (2010), 117–144.
- [76] Serdar Kadioglu and Meinolf Sellmann. 2010. Grammar constraints. *Constraints* 15, 1 (2010), 117–144. <https://doi.org/10.1007/s10601-009-9073-4>
- [77] Juhani Karhumäki, Filippo Mignosi, and Wojciech Plandowski. 2000. The expressibility of languages and relations by word equations. *J. ACM* 47, 3 (2000), 483–505.
- [78] Scott Kausler and Elena Sherman. 2014. Evaluation of String Constraint Solvers in the Context of Symbolic Execution. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*. Association for Computing Machinery, New York, NY, USA, 259–270.
- [79] Adam Kieźun, Vijay Ganesh, Shay Artzi, Philip J. Guo, Pieter Hooimeijer, and Michael D. Ernst. 2012. HAMPI: A Solver for Word Equations over Strings, Regular Expressions, and Context-Free Grammars. *ACM Trans. Software Engineering and Methodology* 21, 4 (2012), article 25.
- [80] Adam Kieźun, Vijay Ganesh, Philip J. Guo, Pieter Hooimeijer, and Michael D. Ernst. 2009. HAMPI: A Solver for String Constraints. In *Proc. 18th Int. Symp. Software Testing and Analysis*. ACM, 105–116.
- [81] James C. King. 1976. Symbolic Execution and Program Testing. *Commun. ACM* 19, 7 (1976), 385–394.
- [82] Nils Klarlund. 1998. Mona & Fido: The logic-automaton connection in practice. In *Computer Science Logic*, Mogens Nielsen and Wolfgang Thomas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 311–326.
- [83] N. Klarlund and M. I. Schwartzbach. 1999. A domain-specific language for regular sets of strings and trees. *IEEE Transactions on Software Engineering* 25, 3 (1999), 378–386.
- [84] Lars Kotthoff. 2016. Algorithm Selection for Combinatorial Search Problems: A Survey. In *Data Mining and Constraint Programming*. LNAI, Vol. 10101. Springer, 149–190.
- [85] Ludwig Krippahl and Pedro Barahona. 2016. Constraining Redundancy to Improve Protein Docking. In *Principles and Practice of Constraint Programming - 22nd International Conference, CP 2016, Toulouse, France, September 5-9, 2016, Proceedings*. 721–732.
- [86] Ludwig Krippahl, Fábio Madeira, and Pedro Barahona. [n.d.]. Constraining Protein Docking with Coevolution Data for Medical Research. In *Artificial Intelligence in Medicine - 14th Conference on Artificial Intelligence in Medicine, AIME 2013, Murcia, Spain, May 29 - June 1, 2013. Proceedings (Lecture Notes in Computer Science)*, Niels Peek, Roque Marín Morales, and Mor Peleg (Eds.), Vol. 7885. Springer, 110–114.
- [87] Guodong Li and Indradeep Ghosh. 2013. PASS: String Solving with Parameterized Array and Interval Automaton. In *Proc. 9th Int. Haifa Verification Conf. (LNCS)*, V. Bertacco and A. Legay (Eds.), Vol. 8244. Springer, 15–31.
- [88] Tianyi Liang, Andrew Reynolds, Cesare Tinelli, Clark Barrett, and Morgan Deters. 2014. A DPLL(T) Theory Solver for a Theory of Strings and Regular Expressions. In *Computer Aided Verification: Proc. 26th Int. Conf. (LNCS)*, A. Biere and R. Bloem (Eds.), Vol. 8559. Springer, 646–662.
- [89] Anthony Widjaja Lin and Pablo Barceló. 2016. String solving with word equations and transducers: towards a logic for analysing mutation XSS. In *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2016, St. Petersburg, FL, USA, January 20 - 22, 2016*, Rastislav Bodik and Rupak Majumdar (Eds.). ACM, 123–136.
- [90] M. H. Lob. 1953. Concatenation as Basis for a Complete System of Arithmetic. *J. Symb. Log.* 18, 1 (1953), 1–6. <https://doi.org/10.2307/2266320>
- [91] Blake Loring, Duncan Mitchell, and Johannes Kinder. 2019. Sound Regular Expression Semantics for Dynamic Symbolic Execution of JavaScript. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019*, Kathryn S. McKinley and Kathleen Fisher (Eds.). ACM, 425–438.
- [92] M. Lothaire. 1997. *Combinatorics on words*. Vol. 17. Cambridge university press.
- [93] M Lothaire. 2005. *Applied combinatorics on words*. Vol. 105. Cambridge University Press.

- [94] Monsieur Lothaire and M Lothaire. 2002. *Algebraic combinatorics on words*. Vol. 90. Cambridge university press.
- [95] Loi Luu, Shweta Shinde, Prateek Saxena, and Brian Demsky. 2014. A model counter for constraints over unbounded strings. In *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '14, Edinburgh, United Kingdom - June 09 - 11, 2014*, Michael F. P. O'Boyle and Keshav Pingali (Eds.). ACM, 565–576.
- [96] Magnus Madsen and Esben Andreasen. 2014. String Analysis for Dynamic Field Access. In *23rd International Conference on Compiler Construction (LNCS)*, Vol. 8409. Springer, 197–217.
- [97] Michael J. Maher. 2009. Open Constraints in a Boundable World. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, 6th International Conference, CPAIOR 2009, Pittsburgh, PA, USA, May 27-31, 2009, Proceedings (Lecture Notes in Computer Science)*, Willem Jan van Hoeve and John N. Hooker (Eds.), Vol. 5547. Springer, 163–177.
- [98] Gennadiy Semenovitch Makanin. 1977. The problem of solvability of equations in a free semigroup. *Matematicheskii Sbornik* 145, 2 (1977), 147–236.
- [99] Laurent D. Michel and Pascal Van Hentenryck. 2012. Constraint Satisfaction over Bit-Vectors. In *Proceedings of the 18th International Conference on Principles and Practice of Constraint Programming (LNCS)*, Vol. 7514. Springer, 527–543.
- [100] Yasuhiko Minamide. 2005. Static approximation of dynamically generated Web pages. In *WWW*. ACM, 432–441.
- [101] Yasuhiko Minamide and Nobuo Otoi. [n.d.]. PHP String Analyzer. Available at <https://sv.c.titech.ac.jp/minamide/phpsa/>.
- [102] Jean Néraud. 2018. Conferences WORDS, years 1997-2017: Open Problems and Conjectures. *Bulletin of the EATCS* 125 (2018). <http://eatcs.org/beatcs/index.php/beatcs/article/view/540>
- [103] Nicholas Nethercote, Peter J. Stuckey, Ralph Becket, Sebastian Brand, Gregory J. Duck, and Guido Tack. 2007. MiniZinc: Towards a Standard CP Modelling Language. In *Proceedings of the 13th International Conference on Principles and Practice of Constraint Programming (LNCS)*, Vol. 4741. Springer, 529–543.
- [104] Changhee Park, Hyeonseung Im, and Sukyoung Ryu. 2016. Precise and scalable static analysis of jQuery using a regular expression domain. In *Proceedings of the 12th Symposium on Dynamic Languages, DLS 2016, Amsterdam, The Netherlands, November 1, 2016*, Roberto Ierusalimsky (Ed.). ACM, 25–36.
- [105] G. Pesant. 2004. A Regular Language Membership Constraint for Finite Sequences of Variables. In *Proceedings of the 10th International Conference on Principles and Practice of Constraint Programming (LNCS)*, M. Wallace (Ed.), Vol. 3258. Springer-Verlag, 482–495.
- [106] Gilles Pesant, Claude-Guy Quimper, Louis-Martin Rousseau, and Meinolf Sellmann. 2009. The Polytope of Context-Free Grammar Constraints. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, 6th International Conference, CPAIOR 2009, Pittsburgh, PA, USA, May 27-31, 2009, Proceedings (Lecture Notes in Computer Science)*, Willem Jan van Hoeve and John N. Hooker (Eds.), Vol. 5547. Springer, 223–232.
- [107] Wojciech Plandowski. 2004. Satisfiability of word equations with constants is in PSPACE. *J. ACM* 51, 3 (2004), 483–496.
- [108] Claude-Guy Quimper and Toby Walsh. 2006. Global Grammar Constraints. In *Principles and Practice of Constraint Programming - CP 2006, 12th International Conference, CP 2006, Nantes, France, September 25-29, 2006, Proceedings (Lecture Notes in Computer Science)*, Frédéric Benhamou (Ed.), Vol. 4204. Springer, 751–755.
- [109] Willard Van Orman Quine. 1946. Concatenation as a Basis for Arithmetic. *J. Symb. Log.* 11, 4 (1946), 105–114. <https://doi.org/10.2307/2268308>
- [110] Arcot Rajasekar. 1994. Applications in constraint logic programming with strings. In *Principles and Practice of Constraint Programming*, Alan Borning (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 109–122.
- [111] Andrew Reynolds, Andres Nötzli, Clark Barrett, and Cesare Tinelli. 2019. High-Level Abstractions for Simplifying Extended String Constraints in SMT. In *Computer Aided Verification*, Isil Dillig and Serdar Tasiran (Eds.). Springer International Publishing, Cham, 23–42.
- [112] Andrew Reynolds, Maverick Woo, Clark Barrett, David Brumley, Tianyi Liang, and Cesare Tinelli. 2017. Scaling Up DPLL(T) String Solvers Using Context-Dependent Simplification. In *Computer Aided Verification*, Rupak Majumdar and Viktor Kunčak (Eds.). Springer International Publishing, Cham, 453–474.
- [113] F. Rossi, P. van Beek, and T. Walsh (Eds.). 2006. *Handbook of Constraint Programming*. Elsevier.
- [114] Arto Salomaa and Ian N. Sneddon. 1969. *Theory of Automata*. Pergamon Press Reprint.
- [115] H. Samimi, M. SchÄdfer, S. Artzi, T. Millstein, F. Tip, and L. Hendren. 2012. Automated repair of HTML generation errors in PHP applications using string constraint solving. In *2012 34th International Conference on Software Engineering (ICSE)*. 277–287.
- [116] Prateek Saxena and Devdatta Akhawe. [n.d.]. Kaluza String Solver. Available at <http://webblaze.cs.berkeley.edu/2010/kaluza/>.
- [117] Prateek Saxena, Devdatta Akhawe, Steve Hanna, Feng Mao, Stephen McCamant, and Dawn Song. 2010. A Symbolic Execution Framework for JavaScript. In *Proc. 2010 IEEE Symp. Security and Privacy*. IEEE Comp. Soc., 513–528.
- [118] Joseph Scott. [n.d.]. Prototype implementation of a bounded string module for the Gecode CP library. Available at <https://github.com/jossc/gecode-string>.
- [119] Joseph D. Scott. 2016. *Other Things Besides Number: Abstraction, Constraint Propagation, and String Variable Types*. Ph.D. Dissertation. Department of Information Technology, Uppsala University, Sweden. Available at <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-273311>.
- [120] Joseph D. Scott, Pierre Flener, and Justin Pearson. 2013. Bounded Strings for Constraint Programming. In *25th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2013, Herndon, VA, USA, November 4-6, 2013*. IEEE Computer Society, 1036–1043.
- [121] Joseph D. Scott, Pierre Flener, and Justin Pearson. 2013. Bounded Strings for Constraint Programming. In *ICTAI*. IEEE Computer Society, 1036–1043.

- [122] Joseph D. Scott, Pierre Flener, and Justin Pearson. 2015. Constraint Solving on Bounded String Variables. In *Twelfth International Conference on Integration of Artificial Intelligence and Operations Research techniques in Constraint Programming (LNCS)*, Vol. 9075. Springer, 375–392.
- [123] Joseph D. Scott, Pierre Flener, Justin Pearson, and Christian Schulte. 2017. Design and Implementation of Bounded-Length Sequence Variables. In *Proc. 14th Int. Conf. Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (LNCS)*, M. Lombardi and D. Salvagnin (Eds.), Vol. 10335. Springer, 51–67.
- [124] David B. Searls. 1995. String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA. *J. Log. Program.* 24, 1&2 (1995), 73–102.
- [125] Jari Stenman. [n.d.]. Norn, a solver for string constraints. Available at <http://user.it.uu.se/~jarst116/norn/>.
- [126] Sanu Subramanian, Murphy Berzish, Yunhui Zheng, Omer Tripp, and Vijay Ganesh. 2016. A Solver for a Theory of Strings and Bit-vectors. *CoRR* abs/1605.09446 (2016). <http://arxiv.org/abs/1605.09446>
- [127] Takaaki Tateishi, Marco Pistoia, and Omer Tripp. 2013. Path- and Index-Sensitive String Analysis Based on Monadic Second-Order Logic. *ACM Trans. Software Engineering Methodology* 22, 4 (2013), article 33.
- [128] BRICS team. [n.d.]. Java String Analyzer. Available at <https://www.brics.dk/JSA>.
- [129] BRICS team. [n.d.]. The MONA Project. Available at <https://www.brics.dk/mona>.
- [130] CVC4 team. [n.d.]. Strings in CVC4. Available at <http://cvc4.cs.stanford.edu/wiki/Strings>.
- [131] Z3 team. [n.d.]. Strings and Sequences in Z3. Available at <https://rise4fun.com/z3/tutorialcontent/sequences>.
- [132] Julian Thomé, Lwin Khin Shar, Domenico Bianculli, and Lionel C. Briand. 2017. Search-driven string constraint solving for vulnerability detection. In *Proceedings of the 39th International Conference on Software Engineering, ICSE 2017, Buenos Aires, Argentina, May 20-28, 2017*. 198–208.
- [133] Emina Torlak and Daniel Jackson. 2007. Kodkod: A Relational Model Finder. In *Tools and Algorithms for the Construction and Analysis of Systems, 13th International Conference, TACAS 2007, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2007 Braga, Portugal, March 24 - April 1, 2007, Proceedings (Lecture Notes in Computer Science)*, Orna Grumberg and Michael Huth (Eds.), Vol. 4424. Springer, 632–647.
- [134] Minh-Thai Trinh, Duc-Hiep Chu, and Joxan Jaffar. 2014. S3: A Symbolic String Solver for Vulnerability Detection in Web Applications. In *SIGSAC*. ACM, 1232–1243.
- [135] Minh-Thai Trinh, Duc-Hiep Chu, and Joxan Jaffar. 2016. Progressive Reasoning over Recursively-Defined Strings. In *Computer Aided Verification - 28th International Conference, CAV 2016, Toronto, ON, Canada, July 17-23, 2016, Proceedings, Part I (Lecture Notes in Computer Science)*, Swarat Chaudhuri and Azadeh Farzan (Eds.), Vol. 9779. Springer, 218–240.
- [136] Minh-Thai Trinh, Duc-Hiep Chu, and Joxan Jaffar. 2017. Model Counting for Recursively-Defined Strings. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II*. 399–418.
- [137] Min-Thai Trinh. [n.d.]. S3: An Efficient String Solver and Model Counter. Available at <https://trinhmt.github.io/home/S3/>.
- [138] Ju M Vazenin and B V Rozenblat. 1983. DECIDABILITY OF THE POSITIVE THEORY OF A FREE COUNTABLY GENERATED SEMIGROUP. *Mathematics of the USSR-Sbornik* 44, 1 (feb 1983), 109–116. <https://doi.org/10.1070/sm1983v04n01abeh000954>
- [139] Margus Veanes and Peli de Halleux. [n.d.]. Rex - Regular Expression Exploration. Available at <https://rise4fun.com/rex>.
- [140] Margus Veanes, Peli de Halleux, and Nikolai Tillmann. 2010. Rex: Symbolic Regular Expression Explorer. In *Third International Conference on Software Testing, Verification and Validation, ICST 2010, Paris, France, April 7-9, 2010*. IEEE Computer Society, 498–507.
- [141] Pavol Voda. 1988. The constraint language trilogy: Semantics and computations. Technical report, Complete Logic Systems, North Vancouver, BC, Canada.
- [142] J. von Neumann. 1951. The general and logical theory of automata. In *Cerebral Mechanisms in Behaviour*, L. A. Jeffress (Ed.). Wiley.
- [143] Clifford Walinsky. 1989. CLP(Sigma*): Constraint Logic Programming with Regular Sets. In *Logic Programming, Proceedings of the Sixth International Conference, Lisbon, Portugal, June 19-23, 1989*, Giorgio Levi and Maurizio Martelli (Eds.). MIT Press, 181–196.
- [144] Hung-En Wang, Tzung-Lin Tsai, Chun-Han Lin, Fang Yu, and Jie-Hong R. Jiang. 2016. String Analysis via Automata Manipulation with Logic Circuit Representation. In *Computer Aided Verification - 28th International Conference, CAV 2016, Toronto, ON, Canada, July 17-23, 2016, Proceedings, Part I (Lecture Notes in Computer Science)*, Swarat Chaudhuri and Azadeh Farzan (Eds.), Vol. 9779. Springer, 241–260.
- [145] Fang Yu, Muath Alkhalaf, and Tefik Bultan. 2010. Stranger: An Automata-Based String Analysis Tool for PHP. In *TACAS (LNCS)*, Vol. 6015. Springer, 154–157.
- [146] Yunhui Zheng, Vijay Ganesh, Sanu Subramanian, Omer Tripp, Murphy Berzish, Julian Dolby, and Xiangyu Zhang. 2017. Z3str2: an efficient solver for strings, regular expressions, and length constraints. *Formal Methods in System Design* 50, 2-3 (2017), 249–288.
- [147] Yunhui Zheng, Vijay Ganesh, Sanu Subramanian, Omer Tripp, Julian Dolby, and Xiangyu Zhang. 2015. Effective Search-Space Pruning for Solvers of String Equations, Regular Expressions and Length Constraints. In *CAV (LNCS)*, Vol. 9206. Springer, 235–254.
- [148] Yunhui Zheng, Xiangyu Zhang, and Vijay Ganesh. 2013. Z3-str: A Z3-Based String Solver for Web Application Analysis. In *Proc. 9th Joint Meeting on Foundations of Software Engineering*. ACM, 114–124.
- [149] Qizhen Zhu, Hitoshi Akama, and Yasuhiko Minamide. 2019. Solving String Constraints with Streaming String Transducers. *JIP* 27 (2019), 810–821.