

Statically Typed String Sanitation Inside a Python

Nathan Fulton

Cyrus Omar

Jonathan Aldrich

Carnegie Mellon University
Pittsburgh, PA

{nathanfu, comar, aldrich}@cs.cmu.edu

ABSTRACT

Web applications ultimately generate strings containing commands, which are executed by systems like web browsers and database engines. Strings constructed from user input that has not been properly sanitized can thus cause command injection vulnerabilities.

In this paper, we introduce *regular string types*, which classify strings known statically to be in a specified regular language and support operations like concatenation, substitution and coercion. Regular string types can be used to implement, in essentially a conventional manner, the parts of a web application or application framework that construct such command strings. Straightforward type annotations at key interfaces can be used to statically verify that sanitization has been performed correctly without introducing redundant run-time checks. We specify this type system as a minimal typed lambda calculus, λ_{RS} .

To be practical, adopting a type system like this should not require adopting a new programming language. Instead, we favor extensible type systems: new static type systems like this should be distributed as libraries atop a mechanism that guarantees that they can be safely composed. We support this by 1) specifying a translation from λ_{RS} to a language containing only strings and regular expressions, then, taking Python as such a language, 2) implement the type system together with the translation as a library using `atlang`, an extensible static type system for Python (being developed by the authors).

1. INTRODUCTION

Command injection vulnerabilities are among the most common and severe in modern web applications [9]. They arise because web applications, at their boundaries, must control external systems that expose string-based command interfaces. For example, web browsers are controlled using HTML and Javascript sent from a server as a string, and database engines execute SQL queries also sent as strings. When these commands contain data derived from user input, care must be taken to ensure that the user cannot provide an

input that will subvert the intended command. For example, a SQL query constructed using string concatenation exposes a SQL injection vulnerability if `name` is controlled by a user:

```
'SELECT * FROM users WHERE name=' + name + ''
```

If a malicious user enters the name `""; DROP TABLE users --`, the entire database could be erased.

To avoid this problem, the program must *sanitize* user input. For example, in this case, the developer (or, more often, a framework) might define a function `sanitize` that prepends double quotes and existing backslashes with a backslash, which SQL treats safely. Note that this function is not idempotent, so it should only be called once. Guaranteeing that user input has already been sanitized before it is used to construct a command is challenging.

We observe that most such sanitization techniques can be understood in terms of *regular languages*. For example, `name` should be a string in the language described by the regular expression `([^\]|(\)|\\))*` – a sequence consisting of characters other than quotation marks or backslashes, or escaped quotation marks or escaped backslashes. Concrete syntax like this can be understood to desugar, in a standard way, to the abstract syntax for regular expressions shown in Figure 1. We will work with this “core” for simplicity, and assume basic familiarity with regular expressions [6].

In this paper, we present a static type system that tracks the regular language a string belongs to. We take advantage of closure and decidability properties of regular languages to support a number of useful operations on values of such *regular string types*. These make it possible to implement sanitation protocols like the one just described in an essentially conventional manner. The result is a system where the fact that a string has been *correctly* sanitized becomes manifest in its type. Missing calls to sanitization functions are detected statically, and, importantly, so are *incorrectly implemented sanitization functions* (i.e. these functions need not be trusted). Run-time checks are only used when going from less precise to more precise types (e.g. at the edges of the system, where user input has not yet been validated) and no additional space overhead is required.

Our type system for regular string types, λ_S , is defined as an extension to a simply-typed lambda calculus. Operations on strings are given types according to corresponding, type-level computations on regular expressions. A concatenation operation is defined in order to demonstrate this principle, but the premier operation of our system is *replacement*, which captures a rich set of input sanitization techniques including escaped characters and encodings. In addition to defining this system, we present a correctness-preserving

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

compilation of λ_S into an underlying language with ordinary strings as a regular expression library.

In addition to this theoretical contribution, we also explain how our extension may be implemented in the Ace Programming Language, which is itself an extension of Python – a popular web programming language.

Summarily, we present a simple type system extension which ensures the absence of input sanitation vulnerabilities by statically checking the correctness and correct use of input sanitation algorithms. This approach is *composable* in the sense that it is a conservative extension. This approach is *complementary* to existing input sanitation techniques that use regular expressions for input sanitation.

1.1 Related Work and Alternative Approaches

The input sanitation problem is well-understood. There exist a large number of techniques and technologies, proposed by both practitioners and researchers, for preventing injection-style attacks. In this section, we explain how our approach to the input sanitation problem differs from each of these approaches. More important than these differences, however, is our more general assertion that language extensibility is a promising approach toward consideration of security goals in programming language design.

Unlike *frameworks and libraries* provided by languages such as Haskell and Ruby, our type system provides a *static* guarantee that input is always properly sanitized before use. Doing so requires reasoning about the operations on regular languages corresponding to standard operations on strings; we are unaware of any production system which contains this form of reasoning. Therefore, even where frameworks and libraries provide a viable interface or wrapper around input sanitation, our approach is complementary because it ensures the correctness of the framework or library itself. Furthermore, our approach is more general than database abstraction layers because our mechanism is applicable to all forms of command injection (e.g. shell injection or remote file inclusion).

A number of research languages provide static guarantees that a program is free of input sanitation vulnerabilities [3]. Unlike this work, our solution to the input sanitation problem has a very low barrier to adoption; for instance, our implementation conservatively extends Python – a popular language among web developers. We also believe our general approach is better-positioned for security, where continuously evolving threats might require frequent addition of new analyses; in these cases, the composability and generality of our approach is a substantial advantage.

The Wyvern programming language provides a general framework for composing language extensions [1][2]. Our work identifies one particular extension, and is therefore complementary to Wyvern and related work on extensible programming languages. We are also unaware of any extensible programming languages which emphasize applications to security concerns.

Incorporating regular expressions into the type system is not novel. The XDuce system [8, 7] checks XML documents against schemas using regular expressions. Similarly, XHaskell [10] focuses on XML documents. We differ from this and related work in at least three ways:

- Our system is defined within an extensible type system.

$$r ::= \epsilon \mid \cdot \mid a \mid r \cdot r \mid r + r \mid r^* \quad a \in \Sigma$$

Figure 1: Regular expressions over the alphabet Σ .

- We demonstrate that regular expression types are applicable to the web security domain, whereas previous work on regular expression types focused on XML schemas.
- Although our static replacement operation is definable in some languages with regular expression types, we are the first to expose this operation and connect the semantics of regular language replacement with the semantics of string substitution via a type safety and compilation correctness argument.

In conclusion, our contribution is a type system, implemented within an extensible type system, for checking the correctness of input sanitation algorithms.

2. A TYPE SYSTEM FOR STRING SANITATION

In this section we define a language for statically checked string sanitation (λ_S). The system has regular expression types `stringin[r]` where r is a regular expression. Expressions of this type evaluate to string literals in the language described by r . Operations on expressions of type `stringin[r]` preserve this property.

The premier operation for manipulating strings in λ_S is string substitution, which is a familiar operation to any programmer who has used regular expressions. The replacement operation replaces all instances of a pattern in one string with another string; for instance, `lsubst(a|b, a, c) = c`. In order to compute the type resulting from substitution, we also need to compute the result of replacing one language with another inside a given language. Finally, just for convenience, we provide a coerce operation. The introduction of coercion requires handling of runtime errors.

The underlying language λ_P has only one type for strings. We prove that whenever a term is translated from λ_S to λ_P , correctness is preserved. The only exception is in the case of unsafe casts in λ_S , which are unnecessary but are included to demonstrate that the regex library of λ_P may be used to insert dynamic checks whenever even when developers are not careful about using statically checked operations.

A brief outline of this section follows:

- Page 3 contains a definition of λ_S , λ_P and the translation from λ_S to λ_P .
- In §2.1 we state some properties about regular expressions which are needed in our correctness proofs.
- In §2.2 we prove type safety for λ_P as well as both type safety and correctness for λ_S .
- In §2.3 we prove that translation preserves the correctness result about λ_S .

$\psi ::= \psi \rightarrow \psi$ $\text{stringin}[r]$	source types
$S ::= \lambda x.e$ ee $\text{rstr}[s]$ $\text{rconcat}(S, S)$ $\text{rreplace}[r](S, S)$ $\text{rcoerce}[r](S)$	source terms $s \in \Sigma^*$

Figure 2: Syntax for the string sanitation fragment of our source language, λ_S .

$\theta ::= \theta \rightarrow \theta$ string regex	target types
$P ::= \lambda x.e$ ee $\text{str}[s]$ $\text{rx}[r]$ $\text{concat}(P, P)$ $\text{preplace}(P, P, P)$ $\text{check}(P, P)$	target terms

Figure 3: Syntax for the fragment of our target language, λ_P , containing strings and statically constructed regular expressions.

$\Psi \vdash S : \psi$	$\Psi ::= \emptyset \mid \Psi, x : \psi$
$\frac{\text{S-T-STRINGIN-I} \quad s \in \mathcal{L}\{r\}}{\Psi \vdash \text{rstr}[s] : \text{stringin}[r]}$	
$\frac{\text{S-T-CONCAT} \quad \Psi \vdash S_1 : \text{stringin}[r_1] \quad \Psi \vdash S_2 : \text{stringin}[r_2]}{\Psi \vdash \text{rconcat}(S_1, S_2) : \text{stringin}[r_1 \cdot r_2]}$	
$\frac{\text{S-T-REPLACE} \quad \Psi \vdash S_1 : \text{stringin}[r_1] \quad \Psi \vdash S_2 : \text{stringin}[r_2] \quad \text{lreplace}(r, r_1, r_2) = r'}{\Psi \vdash \text{rreplace}[r](S_1, S_2) : \text{stringin}[r']}$	
$\frac{\text{S-T-COERCE} \quad \Psi \vdash S : \text{stringin}[r']}{\Psi \vdash \text{rcoerce}[r](S) : \text{stringin}[r]}$	

Figure 4: Typing rules for our fragment of λ_S . The typing context Ψ is standard.

$S \Downarrow S$	$S \text{ err}$
$\frac{\text{S-E-RSTR} \quad \text{rstr}[s] \Downarrow \text{rstr}[s]}{\text{rstr}[s] \Downarrow \text{rstr}[s]}$	
$\frac{\text{S-E-CONCAT} \quad S_1 \Downarrow \text{rstr}[s_1] \quad S_2 \Downarrow \text{rstr}[s_2]}{\text{rconcat}(S_1, S_2) \Downarrow \text{rstr}[s_1 s_2]}$	
$\frac{\text{S-E-REPLACE} \quad S_1 \Downarrow \text{rstr}[s_1] \quad S_2 \Downarrow \text{rstr}[s_2] \quad \text{lsubst}(r, s_1, s_2) = s}{\text{rreplace}[r](S_1, S_2) \Downarrow \text{rstr}[s]}$	
$\frac{\text{S-E-COERCE-OK} \quad S \Downarrow \text{rstr}[s] \quad s \in \mathcal{L}\{r\}}{\text{rcoerce}[r](S) \Downarrow \text{rstr}[s]}$	
$\frac{\text{S-E-COERCE-ERR} \quad S \Downarrow \text{rstr}[s] \quad s \notin \mathcal{L}\{r\}}{\text{rcoerce}[r](S) \text{ err}}$	

Figure 5: Big step semantics for our fragment of λ_S . Error propagation rules are omitted.

$$\boxed{\llbracket S \rrbracket = P}$$

$$\begin{array}{c}
\text{Tr-STRING} \\
\frac{}{\llbracket \text{rstr}[s] \rrbracket = \text{str}[s]}
\end{array}
\quad
\begin{array}{c}
\text{Tr-CONCAT} \\
\frac{\llbracket S_1 \rrbracket = P_1 \quad \llbracket S_2 \rrbracket = P_2}{\llbracket \text{rconcat}(S_1, S_2) \rrbracket = \text{concat}(P_1, P_2)}
\end{array}
\quad
\begin{array}{c}
\text{Tr-SUBST} \\
\frac{\llbracket S_1 \rrbracket = P_1 \quad \llbracket S_2 \rrbracket = P_2}{\llbracket \text{rreplace}[r](S_1, S_2) \rrbracket = \text{replace}(\text{rx}[r], P_1, P_2)}
\end{array}$$

$$\begin{array}{c}
\text{Tr-COERCE-OK} \\
\frac{S : \text{rstr}[r] \quad \mathcal{L}\{r'\} \subseteq \mathcal{L}\{r\}}{\llbracket \text{rcoerce}[r'](S) \rrbracket = \text{str}[s]}
\end{array}
\quad
\begin{array}{c}
\text{Tr-COERCE-NOTOK} \\
\frac{\llbracket S \rrbracket = P \quad S : \text{rstr}[r] \quad \mathcal{L}\{r'\} \not\subseteq \mathcal{L}\{r\}}{\llbracket \text{rcoerce}[r'](S) \rrbracket = \text{check}(\text{rx}[r'], P)}
\end{array}$$

Figure 8: Translation from source terms (S) to target terms (P). The translation is type-directed in the Tr-Coerce cases.

$$\boxed{\Theta \vdash P : \theta} \quad \Theta ::= \emptyset \mid \Theta, x : \theta$$

$$\begin{array}{c}
\text{P-T-STRING} \\
\frac{}{\Theta \vdash \text{str}[s] : \text{string}}
\end{array}
\quad
\begin{array}{c}
\text{P-T-REGEX} \\
\frac{}{\Theta \vdash \text{rx}[r] : \text{regex}}
\end{array}$$

$$\begin{array}{c}
\text{P-T-CONCAT} \\
\frac{\Theta \vdash P_1 : \text{string} \quad \Theta \vdash P_2 : \text{string}}{\Theta \vdash \text{concat}(P_1, P_2) : \text{string}}
\end{array}$$

$$\begin{array}{c}
\text{P-T-REPLACE} \\
\frac{\Theta \vdash P_1 : \text{regex} \quad \Theta \vdash P_2 : \text{string} \quad \Theta \vdash P_3 : \text{string}}{\Theta \vdash \text{preplace}(P_1, P_2, P_3) : \text{string}}
\end{array}$$

$$\begin{array}{c}
\text{P-T-CHECK} \\
\frac{\Theta \vdash P_1 : \text{regex} \quad \Theta \vdash P_2 : \text{string}}{\Theta \vdash \text{check}(P_1, P_2) : \text{string}}
\end{array}$$

Figure 6: Typing rules for our fragment of λ_P . The typing context Θ is standard.

$$\boxed{P \Downarrow P} \quad \boxed{P \text{ err}}$$

$$\begin{array}{c}
\text{P-E-STR} \\
\frac{}{\text{str}[s] \Downarrow \text{str}[s]}
\end{array}
\quad
\begin{array}{c}
\text{P-E-RX} \\
\frac{}{\text{rx}[r] \Downarrow \text{rx}[r]}
\end{array}
\quad
\begin{array}{c}
\text{P-E-CONCAT} \\
\frac{P_1 \Downarrow \text{str}[s_1] \quad P_2 \Downarrow \text{str}[s_2]}{\text{concat}(P_1, P_2) \Downarrow \text{str}[s_1 s_2]}
\end{array}$$

$$\begin{array}{c}
\text{P-E-REPLACE} \\
\frac{P_2 \Downarrow \text{str}[s_2] \quad P_3 \Downarrow \text{str}[s_3] \quad \text{lsubst}(r, s_2, s_3) = s}{\text{preplace}(P_1, P_2, P_3) \Downarrow \text{str}[s]}
\end{array}$$

$$\begin{array}{c}
\text{P-E-CHECK-OK} \\
\frac{P_1 \Downarrow \text{rx}[r] \quad P_2 \Downarrow \text{rstr}[s] \quad s \in \mathcal{L}\{r\}}{\text{check}(P_1, P_2) \Downarrow \text{str}[s]}
\end{array}$$

$$\begin{array}{c}
\text{P-E-CHECK-ERR} \\
\frac{P_1 \Downarrow \text{rx}[r] \quad P_2 \Downarrow \text{str}[s] \quad s \notin \mathcal{L}\{r\}}{\text{check}(P_1, P_2) \text{ err}}
\end{array}$$

Figure 7: Big step semantics for our fragment of λ_P . Error propagation rules are omitted.

2.1 Properties of Regular Languages

Our type safety proof for language S relies on a relationship between string substitution and language substitution given in lemma 5. We also rely upon several other properties of regular languages. Throughout this section, we fix an alphabet Σ over which strings s and regular expressions r are defined. throughout the paper, $\mathcal{L}\{r\}$ refers to the language recognized by the regular expression r . This distinction between the regular expression and its language – typically elided in the literature – makes our definition and proofs about systems S and P more readable.

Lemma 1. *Properties of Regular Languages and Expressions. The following are properties of regular expressions which are necessary for our proofs: If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then $s_1s_2 \in \mathcal{L}\{r_1r_2\}$. For all strings s and regular expressions r , either $s \in \mathcal{L}\{r\}$ or $s \notin \mathcal{L}\{r\}$. Regular languages are closed under difference, right quotient, reversal, and string homomorphism.*

If any of these properties are unfamiliar, the reader may refer to a standard text on the subject [6].

Definition 2 (lsubst). The replation $\text{lsubst}(r, s_1, s_2) = s$ produces a string s in which all substrings of s_1 matching r are replaced with s_2 .

Definition 3 (lreplace). The relation $\text{lreplace}(r, r_1, r_2) = r'$ relates r, r_1 , and r_2 to a language r' containing all strings of r_1 except that any substring $s_{pre}ss_{post} \in \mathcal{L}\{r_1\}$ where $s \in \mathcal{L}\{r\}$ is replaced by the set of strings $s_{pre}s_2s_{post}$ for all $s_2 \in \mathcal{L}\{r_2\}$ (the prefix and postfix positions may be empty).

Lemma 4. Closure. *If $\mathcal{L}\{r\}, \mathcal{L}\{r_1\}$ and $\mathcal{L}\{r_2\}$ are regular expressions, then $\mathcal{L}\{\text{lreplace}(r, r_1, r_2)\}$ is also a regular language.*

Proof. The theorem follows from closure under difference, right quotient, reversal and string homomorphism. \square

Lemma 5. Substitution Correspondence. *If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then $\text{lsubst}(r, s_1, s_2) \in \mathcal{L}\{\text{lreplace}(r, s_1, s_2)\}$.*

Proof. The theorem follows from the definitions of lsubst and lreplace; note that language substitutions over-approximate string substitutions. \square

2.2 Safety of the Source and Target Languages

Lemma 6. *If $\Psi \vdash S : \text{stringin}[r]$ then r is a well-formed regular expression.*

Proof. The only non-trivial case is S-T-Replace, which follows from lemma 4. \square

Lemma 7. *If $\Theta \vdash P : \text{regex}$ then $P \Downarrow \text{rx}[r]$ such that r is a well-formed regular expression.*

We now prove safety for the string fragment of the source and target languages.

Theorem 8. Safety and Sanitation Correctness for the String Fragment of P . *Let S be a term in the source language. If $\Psi \vdash S : \text{stringin}[r]$ then $S \Downarrow \text{rstr}[s]$ and $\Psi \vdash \text{rstr}[s] : \text{stringin}[r]$; or else $S \text{ err}$.*

Proof. By induction on the typing relation, where (a) case holds by lemma 1 in the S-T-Concat case and lemma 5 in the S-T-Replace case.. The (b) cases hold by unstated, but standard, error propagation rules. \square

In addition to safety, we proof a correctness result for λ_S which ensures that well-typed terms of regular string type are in the language associated with their type.

Theorem 9. Correctness of Input Sanitation for λ_S . *If $\Psi \vdash S : \text{stringin}[r]$ and $S \Downarrow \text{rstr}[s]$ then $s \in \mathcal{L}\{r\}$.*

Proof. Follows directly from type safety, canonical forms for λ_S . \square

Theorem 10. *Let P be a term in the target language. If $\Theta \vdash P : \theta$ then $P \Downarrow P'$ and $\Theta \vdash P' : \theta$, or else $P \text{ err}$.*

2.3 Translation Correctness

The main theorem of this paper is Theorem 12. Establishing this result requires an additional theorem (Theorem ??), which establishes a relationship between canonical forms for the string fragments of λ_S and λ_P .

Theorem 11. Translation Correctness. *If $\Psi \vdash S : \text{stringin}[r]$ then there exists a P such that $\llbracket S \rrbracket = P$ and either: (a) $P \Downarrow \text{str}[s]$ and $S \Downarrow \text{rstr}[s]$, or (b) $P \text{ err}$ and $S \text{ err}$.*

Proof. The proof proceeds by induction on the typing relation for S and an appropriate choice of P ; in each case, the choice is obvious. The subcases (a) proceed by inversion and appeals to our type safety theorems as well as the induction hypothesis. The subcases (b) proceed by the standard error propagation rules omitted for space. Throughout the proof, properties from the closure lemma for regular languages are necessary. \square

Finally, our main theorem establishes that input sanitation correctness of λ_S is preserved under the translation into λ_P .

Theorem 12. Correctness of Input Sanitation for Translated Terms. *If $\llbracket S \rrbracket = P$ and $\Psi \vdash S : \text{stringin}[r]$ then either $P \text{ err}$ or $P \Downarrow \text{str}[s]$ for $s \in \mathcal{L}\{r\}$.*

Proof. By theorem 11, $P \Downarrow \text{str}[s]$ implies that $S \Downarrow \text{rstr}[s]$. By theorem 9, this together with the assumption that S is well-typed implies that $s \in \mathcal{L}\{r\}$. \square

3. IMPLEMENTATION IN ATLANG

www Here, we use the argument annotation syntax introduced in Python 3 (a similar syntax is available for Python 2.6+, not shown).

In this section, we present an example atlang program using our extension together with our implementation of the extension in Atlang modulo lreplace. Note that in addition to Figure 1, regular expressions in atlang also support complementation using the carot.

TODO-nrf insert a paragraph or two which explains what's going on...

```

1 #todo: why are these r's here?
2 #TODO-nrf this is maybe kind-of right bue I think there's a nice
3 @fn
4 def sanitize(s : string_in[r'.*']):
5     return (s.replace(r'"', '&quot;')
6             .replace(r'<', '&lt;')
7             .replace(r'>', '&gt;'))
8
9 @fn
10 def results_query(s : string_in[r'^[\"']*']):
11     return 'SELECT * FROM users WHERE name="' + s + '"'
12
13 @fn
14 def results_div(s : string_in[r'<*>.*']):
15     return '<div>Results for ' + s + '</div>'
16
17 def main(db):
18     input = sanitize(user_input())
19     results = db.execute(results_query(input))
20     return results_div(input) + format(results)

```

Figure 8: Regular string types in atlang, a library that enables static type checking for Python.

3.1 An Example

Figure ?? demosntrates the use of string types in atlang. The sanitize function takes an arbitrary string and returns a string without double quotes and left and right brackets. In this example, we use HTML escape sequences; many variants of SQL provide their own syntax for this as well.

TODO-nrf line by line of example.

3.2 Extension Implementation

TODO-nrf line by line of extension implementation as ewll.

An implementation of a similar extension as the one presented here is discussed in [4] and [5]. Our prior work only supports a special case of replacement where unsafe substrings are completely removed. In practice, most libraries and frameworks escape command sequences instead of stripping characters; therefore, we plan to extend our implementation with support for the replace operation as described in this paper.

4. CONCLUSION

Composable analyses which complement existing approaches constitute a promising approach toward the integration of security concerns into programming languages. In this paper, we presented a system with both of these properties and defined a security-preserving transformation. Unlike other approaches, our solution complements existing, familiar solutions while providing a strong guarantee that traditional library and framework-based approaches are implemented and utilized correctly.

5. REFERENCES

- [1] A. Chlipala. Static checking of dynamically-varying security policies in database-backed applications. In *OSDI'10: Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation*, Oct. 2010.
- [2] N. Fulton. Security through extensible type systems. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, pages 107–108, New York, NY, USA, 2012. ACM.

```

1 class string_in(atlang.Type):
2     def __init__(self, rx):
3         rx = rx.normalize(rx)
4         atlang.Type.__init__(idx=rx)
5
6     def ana_Str(self, ctx, node):
7         if not in_lang(node.s, self.idx):
8             raise atlang.TypeError("...", node)
9
10    def trans_Str(self, ctx, node):
11        return astx.copy(node)
12
13    def syn_BinOp_Add(self, ctx, node):
14        left_t = ctx.syn(node.left)
15        right_t = ctx.syn(node.right)
16        if isinstance(left_t, string_in):
17            left_rx = left_t.idx
18            if isinstance(right_t, string_in):
19                right_rx = right_t.idx
20                return string_in[lconcat(left_rx, right_rx)]
21            raise atlang.TypeError("...", node)
22
23    def trans_BinOp_Add(self, ctx, node):
24        return astx.copy(node)
25
26    def syn_Method_replace(self, ctx, node):
27        [rx, exp] = node.args
28        if not isinstance(rx, ast.Str):
29            raise atlang.TypeError("...", node)
30        rx = rx.s
31        exp_t = ctx.syn(exp)
32        if not isinstance(exp_t, string_in):
33            raise atlang.TypeError("...", node)
34        exp_rx = exp_t.idx
35        return string_in[lreplace(self.idx, rx, exp_rx)]
36
37    def trans_Method_replace(self, ctx, node):
38        return astx.quote(
39            """__import__(re); re.sub(%0, %1, %2)""",
40            astx.Str(s=node.args[0]),
41            astx.copy(node.func.value),
42            astx.copy(node.args[1]))
43
44    def syn_Method_check(self, ctx, node):
45        [rx] = node.args
46        if not isinstance(rx, ast.Str):
47            raise atlang.TypeError("...", node)
48        return string_in[rx.s]
49
50    def trans_Method_check(self, ctx, node):
51        return astx.quote(
52            """__import__(string_in_helper);
53            string_in_helper.coerce(%0, %1)""",
54            astx.Str(s=other_t.idx),
55            astx.copy(node))
56
57    def check_Coerce(self, ctx, node, other_t):
58        # coercions can only be defined between
59        # types with the same type constructor,
60        if rx_sublang(other_t.idx, self.idx):
61            return other_t
62        else: raise atlang.TypeError("...", node)

```

Figure 9: Implementation of the `string_in` type constructor in atlang.

```

1 output of successful compilation

```

Figure 10: Output of successful compilation.

```

1 output of failed compilation

```

Figure 11: Output of failed compilation.

- [3] N. Fulton. A typed lambda calculus for input sanitation. Undergraduate thesis in mathematics, Carthage College, 2013.
- [4] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.
- [5] H. Hosoya and B. C. Pierce. XDuce: A statically typed XML processing language. *ACM Transactions on Internet Technology*, 3(2):117–148, May 2003.
- [6] H. Hosoya, J. Vouillon, and B. C. Pierce. Regular Expression Types for XML. In *ICFP '00*, 2000.
- [7] L. Nistor, D. Kurilova, S. Balzer, B. Chung, A. Potanin, and J. Aldrich. Wyvern: A simple, typed, and pure object-oriented language. In *Proceedings of the 5th Workshop on Mechanisms for Specialization, Generalization and Inheritance, MASPEGHI '13*, pages 9–16, New York, NY, USA, 2013. ACM.
- [8] C. Omar, D. Kurilova, L. Nistor, B. Chung, A. Potanin, and J. Aldrich. Safely composable type-specific languages. In R. Jones, editor, *ECOOP 2014 – Object-Oriented Programming*, volume 8586 of *Lecture Notes in Computer Science*, pages 105–130. Springer Berlin Heidelberg, 2014.
- [9] OWASP. Open web application security project top 10.
- [10] M. Sulzmann and K. Lu. Xhaskell – adding regular expression types to haskell. In O. Chitil, Z. Horváth, and V. Zsók, editors, *Implementation and Application of Functional Languages*, volume 5083 of *Lecture Notes in Computer Science*, pages 75–92. Springer Berlin Heidelberg, 2008.