Statically Typed String Sanitation Inside a Python (Technical Report)

Nathan Fulton **Cyrus Omar** **Jonathan Aldrich**

December 2014
CMU-ISR-14-112

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This report contains supporting evidence for claims put forth and explained in the paper "Statically Typed String Sanitation Inside a Python" [1], including proofs of lemmas and theorems asserted in the paper, examples, additional discussion of the paper's technical content, and errata.

This work was supported by the National Security Agency lablet contract #H98230-14-C-0140.

Keywords: type systems; regular languages; input sanitation; string sanitation

Contents

List of Figures

1 Terminology and Notation

Theorems and lemmas appearing in [1] are numbered correspondingly, while supporting facts appearing only in the Technical Report are lettered. Throughout this technical report, we use a small step semantics corresponding to the big step semantics given in [1].

2 Regular Expressions

The syntax of regular expressions over some alphabet Σ is shown in Figure 1.

Assumption A (Regular Expression Congruences)**.** *We assume regular expressions are implicitly identified up to the following congruences:*

$$\begin{aligned}
\epsilon \cdot r &\equiv r \\
r \cdot \epsilon &\equiv r \\
(r_1 \cdot r_2) \cdot r_3 &\equiv r_1 \cdot (r_2 \cdot r_3) \\
r_1 + r_2 &\equiv r_2 + r_1 \\
(r_1 + r_2) + r_3 &\equiv r_1 + (r_2 + r_3) \\
\epsilon^* &\equiv \epsilon
\end{aligned}$$

Assumption B (Properties of Regular Languages)**.** *We assume the following properties:*

1. *If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then $s_1 s_2 \in \mathcal{L}\{r_1 \cdot r_2\}$.*
2. *For all strings s and regular expressions r, either $s \in \mathcal{L}\{r\}$ or $s \notin \mathcal{L}\{r\}$.*
3. *Regular languages are closed under reversal.*

3 λ_{RS}

The syntax of λ_{RS} is specified in Figure 2.

3.1 Static Semantics

The static semantics of λ_{RS} is specified in Figure 4. The typing context obeys the standard structural properties of weakening, exchange and contraction.

3.1.1 Case Analysis

The following correctness conditions must hold for any definition of $\mathsf{lhead}(r)$ and $\mathsf{ltail}(r)$.

Condition C (Correctness of Head)**.** *If $c_1 s' \in \mathcal{L}\{r\}$, then $c_1 \in \mathcal{L}\{\mathsf{lhead}(r)\}$.*

Condition D (Correctness of Tail)**.** *If $c_1 s' \in \mathcal{L}\{r\}$ then $s' \in \mathcal{L}\{\mathsf{ltail}(r)\}$.*

For example, we conjecture (but do not here prove) that the definitions below satisfy these conditions. Note that these are slightly amended relative to the published paper.

Definition 1 (Definition of $\mathsf{lhead}(r)$)**.** We first define an auxiliary relation that determines the set of characters that the head might be, tracking the remainder of any sequences that appear:

$$\begin{aligned}
\mathsf{lhead}(\epsilon, \epsilon) &= \emptyset \\
\mathsf{lhead}(\epsilon, r') &= \mathsf{lhead}(r', \epsilon) \\
\mathsf{lhead}(a, r') &= \{a\} \\
\mathsf{lhead}(r_1 \cdot r_2, r') &= \mathsf{lhead}(r_1, r_2 \cdot r') \\
\mathsf{lhead}(r_1 + r_2, r') &= \mathsf{lhead}(r_1, r') \cup \mathsf{lhead}(r_2, r') \\
\mathsf{lhead}(r^*, r') &= \mathsf{lhead}(r, \epsilon) \cup \mathsf{lhead}(r', \epsilon)
\end{aligned}$$

We define $\mathsf{lhead}(r) = a_1 + a_2 + ... + a_i$ iff $\mathsf{lhead}(r, \epsilon) = \{a_1, a_2, ..., a_i\}$.

Definition 2 (Brzozowski's Derivative)**.** The *derivative of r with respect to s* is denoted by $\delta_s(r)$ and is $\delta_s(r) = \{t | st \in \mathcal{L}\{r\}\}$.

Definition 3 (Definition of $\mathsf{ltail}(r)$)**.** If $\mathsf{lhead}(r, \epsilon) = \{a_1, a_2, ..., a_i\}$, then we define $\mathsf{ltail}(r) = \delta_{a_1}(r) + \delta_{a_2}(r) + ... + \delta_{a_i}(r)$.

3.1.2 Replacement

The following correctness condition must hold for any definition of $\mathsf{lreplace}(r, r_1, r_2)$.

Condition E (Replacement Correctness)**.** *If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then*

$$\mathsf{replace}(r; s_1; s_2) \in \mathcal{L}\{\mathsf{lreplace}(r, r_1, r_2)\}$$

We do not give a particular definition for $\mathsf{lreplace}(r, r_1, r_2)$ here.

3.2 Dynamic Semantics

Figure 5 specifies a small-step operational semantics for λ_{RS}.

3.2.1 Canonical Forms

Lemma F (Canonical Forms)**.** *If $\emptyset \vdash v : \sigma$ then:*

1. *If $\sigma = \mathsf{stringin}[r]$ then $v = \mathsf{rstr}[s]$ and $s \in \mathcal{L}\{r\}$.*
2. *If $\sigma = \sigma_1 \rightarrow \sigma_2$ then $v = \lambda x.e'$.*

Proof. By inspection of the static and dynamic semantics. □

3.2.2 Type Safety

Lemma G (Progress)**.** *If $\emptyset \vdash e : \sigma$ either $e = v$ or $e \mapsto e'$.*

Proof. The proof proceeds by rule induction on the derivation of $\emptyset \vdash e : \sigma$.

λ **fragment**. Cases SS-T-Var, SS-T-Abs, and SS-T-App are exactly as in a proof of progress for the simply typed lambda calculus.

S-T-Stringin-I. Suppose $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[s]$. Then $e = \mathsf{rstr}[s]$.

S-T-Concat. Suppose $\emptyset \vdash \mathsf{rconcat}(e_1; e_2) : \mathsf{stringin}[r_1 \cdot r_2]$ and $\emptyset \vdash e_1 : \mathsf{stringin}[r_1]$ and $\emptyset \vdash e_2 : \mathsf{stringin}[r_2]$. By induction, $e_1 \mapsto e_1'$ or $e_1 = v_1$ and similarly, $e_2 \mapsto e_2'$ or $e_2 = v_2$. If e_1 steps, then SS-E-Concat-Left applies and so $\mathsf{rconcat}(e_1; e_2) \mapsto \mathsf{rconcat}(e_1'; e_2)$. Similarly, if e_2 steps then e steps by SS-E-Concat-Right.

In the remaining case, $e_1 = v_1$ and $e_2 = v_2$. But then it follows by Canonical Forms that $e_1 = \mathsf{rstr}[s_1]$ and $e_2 = \mathsf{rstr}[s_2]$. Finally, by SS-E-Concat, $\mathsf{rconcat}(\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[s_1 s_2]$.

S-T-Case. Suppose $e = \mathsf{rstrcase}(e_1; e_2; x, y.e_3)$ and $\emptyset \vdash e_1 : \mathsf{stringin}[r]$. By induction and Canonical Forms it follows that $e_1 \mapsto e_1'$ or $e_1 = \mathsf{rstr}[s]$. In the former case, e steps by S-E-Case-Left. In the latter case, note that $s = \epsilon$ or $s = at$ for some string t. If $s = \epsilon$ then e steps by S-E-Case-ϵ-Val, and if $s = at$ the e steps by S-E-Case-Concat.

S-T-Replace. Suppose $e = \mathsf{rreplace}[r](e_1; e_2)$, $\emptyset \vdash e : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$ and:

$$\emptyset \vdash e_1 : \mathsf{stringin}[r_1] \tag{1}$$

$$\emptyset \vdash e_2 : \mathsf{stringin}[r_2] \tag{2}$$

By induction on (1), $e_1 \mapsto e_1'$ or $e_1 = v_1$ for some e_1'. If $e_1 \mapsto e_1'$ then e steps by SS-E-Replace-Left. Similarly, if e_2 steps then e steps by SS-E-Replace-Right. The only remaining case is where $e_1 = v_1$ and also $e_2 = v_2$. By Canonical Forms, $e_1 = \mathsf{rstr}[s_1]$ and $e_2 = \mathsf{rstr}[s_2]$. Therefore, $e \mapsto \mathsf{rstr}[\mathsf{replace}(r; s_1; s_2)]$ by SS-E-Replace.

S-T-SafeCoerce. Suppose that $\emptyset \vdash \mathsf{rcoerce}[r](e_1) : \mathsf{stringin}[r]$. and $\emptyset \vdash e_1 : \mathsf{stringin}[r']$ for $\mathcal{L}\{r'\} \subseteq \mathcal{L}\{r\}$. By induction, $e_1 = v_1$ or $e_1 \mapsto e_1'$ for some e_1'. If $e_1 \mapsto e_1'$ then e steps by SS-E-SafeCoerce-Step. Otherwise, $e_1 = v$ and by Canonical Forms $e_1 = \mathsf{rstr}[s]$. In this case, $e = \mathsf{rcoerce}[r](\mathsf{rstr}[s]) \mapsto \mathsf{rstr}[s]$ by SS-E-SafeCoerce.

S-T-Check Suppose that $\emptyset \vdash \mathsf{rcheck}[r](e_0; x.e_1; e_2) : \mathsf{stringin}[r]$ and:

$$\emptyset \vdash e_0 : \mathsf{stringin}[r_0] \tag{3}$$

$$\emptyset, x : \mathsf{stringin}[r] \vdash e_1 : \sigma \tag{4}$$

$$\emptyset \vdash e_2 : \sigma \tag{5}$$

By induction, $e_0 \mapsto e_0'$ or $e_0 = v$. In the former case e steps by SS-E-Check-StepLeft. Otherwise, $e_0 = \mathsf{rstr}[s]$ by Canonical Forms. By Lemma B part 2, either $s \in \mathcal{L}\{r_0\}$ or $s \notin \mathcal{L}\{r_0\}$. In the former case e takes a step by SS-E-Check-Ok. In the latter case e takes a step by SS-E-Check-NotOk.

$\square$

Assumption H (Substitution)**.** *If* $\Psi, x : \sigma' \vdash e : \sigma$ *and* $\Psi \vdash e' : \sigma'$*, then* $\Psi \vdash [e'/x]e : \sigma$.

Lemma I (Preservation for Small Step Semantics)**.** *If* $\emptyset \vdash e : \sigma$ *and* $e \mapsto e'$ *then* $\emptyset \vdash e' : \sigma$.

Proof. By induction on the derivation of $e \mapsto e'$ and $\emptyset \vdash e : \sigma$.

λ **fragment**. Cases SS-E-AppLeft, SS-E-AppRight, and SS-E-AppAbs are exactly as in a proof of type safety for the simply typed lambda calculus.

S-E-Concat-Left. Suppose $e = \mathsf{rconcat}(e_1; e_2) \mapsto \mathsf{rconcat}(e_1'; e_2)$ and $e_1 \mapsto e_1'$. The only rule that applies is S-T-Concat, so $\emptyset \vdash e_1 : \mathsf{stringin}[r_1]$ and $\emptyset \vdash e_2 : \mathsf{stringin}[r_2]$. By induction, $\emptyset \vdash e_1' : \mathsf{stringin}[r_1]$. Therefore, by S-T-Concat, $\emptyset \vdash \mathsf{rconcat}(e_1'; e_2) : \mathsf{stringin}[r_1 r_2]$.

S-E-Concat-Right. Suppose $e = \mathsf{rconcat}(e_1; e_2) \mapsto \mathsf{rconcat}(e_1; e_2')$ and $e_2 \mapsto e_2'$. The only rule that applies is S-T-Concat, so $\emptyset \vdash e_1 : \mathsf{stringin}[r_1]$ and $\emptyset \vdash e_2 : \mathsf{stringin}[r_2]$. By induction, $\emptyset \vdash e_2' : \mathsf{stringin}[r_2]$. Therefore, by S-T-Concat, $\emptyset \vdash \mathsf{rconcat}(e_1; e_2') : \mathsf{stringin}[r_1 r_2]$.

S-E-Concat. Suppose $\mathsf{rconcat}(\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[s_1 s_2]$. The only applicable rule is S-T-Concat, so $\emptyset \vdash \mathsf{rstr}[s_1] : \mathsf{stringin}[r_1]$ and $\emptyset \vdash \mathsf{rstr}[s_2] : \mathsf{stringin}[r_2]$ and $\emptyset \vdash \mathsf{rconcat}(\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) : \mathsf{stringin}[r_1 \cdot r_2]$. By Canonical Forms, $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ from which it follows by Lemma B that $s_1 s_2 \in \mathcal{L}\{r_1 \cdot r_2\}$. Therefore, $\emptyset \vdash \mathsf{rstr}[s_1 s_2] : \mathsf{stringin}[r_1 \cdot r_2]$ by S-T-Rstr.

S-E-Case-Left. Suppose $e \mapsto \mathsf{rstrcase}(e_1'; e_2; x, y.e_3)$ and $\emptyset \vdash e : \sigma$ and $e_1 \mapsto e_1'$. The only rule that applies is S-T-Case, so:

$$\emptyset \vdash e_1 : \mathsf{stringin}[r] \tag{6}$$
$$\emptyset \vdash e_2 : \sigma \tag{7}$$
$$\emptyset, x : \mathsf{stringin}[\mathsf{lhead}(r)], y : \mathsf{stringin}[\mathsf{ltail}(r)] \vdash e_3 : \sigma \tag{8}$$

By (6) and the assumption that $e_1 \mapsto e_1'$, it follows by induction that $\emptyset \vdash e_1' : \mathsf{stringin}[r]$. This fact together with (7) and (8) implies by S-T-Case that $\emptyset \vdash \mathsf{rstrcase}(e_1'; e_2; x, y.e_3) : \sigma$.

S-E-Case-ϵ-Val. Suppose $\mathsf{rstrcase}(e_0; e_2; x, y.e_3) \mapsto e_2$. The only rule that applies is S-T-Case, so $\emptyset \vdash e_2 : \sigma$.

S-E-Case-Concat. Suppose that $e = \mathsf{rstrcase}(\mathsf{rstr}[as]; e_2; x, y.e_3) \mapsto [\mathsf{rstr}[a], \mathsf{rstr}[s]/x, y]e_3$ and that $\emptyset \vdash e : \sigma$. The only rule that applies is S-T-Case so:

$$\emptyset \vdash \mathsf{rstr}[as] : \mathsf{stringin}[r] \tag{9}$$
$$\emptyset \vdash e_2 : \sigma \tag{10}$$
$$\emptyset, x : \mathsf{stringin}[\mathsf{lhead}(r)], y : \mathsf{stringin}[\mathsf{ltail}(r)] \vdash e_3 : \sigma \tag{11}$$

We know that $as \in \mathcal{L}\{r\}$ by Canonical Forms on (9) Therefore, $a \in \mathcal{L}\{\mathsf{lhead}(r)\}$ by Condition C and $s \in \mathcal{L}\{\mathsf{ltail}(r)\}$ by Condition D.

From these facts about a and s we know by S-T-Rstr that $\emptyset \vdash \mathsf{rstr}[a] : \mathsf{stringin}[\mathsf{lhead}(r)]$ and $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[\mathsf{ltail}(r)]$. It follows by Assumption H that $\emptyset \vdash [\mathsf{rstr}[a], \mathsf{rstr}[s]/x, y]e_3 : \sigma$.

Case S-E-Replace-Left. Suppose that $e = \mathsf{rreplace}[r](e_1; e_2) \mapsto \mathsf{rreplace}[r](e_1'; e_2)$ when $e_1 \mapsto e_1'$. The only rule that applies is S-T-Replace, so $\emptyset \vdash e : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$ where:

$$\emptyset \vdash e_1 : \mathsf{stringin}[r_1]$$
$$\emptyset \vdash e_2 : \mathsf{stringin}[r_2]$$

By induction, $\emptyset \vdash e_1' : \mathsf{stringin}[r_1]$. Therefore, $\emptyset \vdash \mathsf{rreplace}[r](e_1'; e_2) : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$ by S-T-Replace.

Case S-E-Replace-Right. Suppose that $e = \mathsf{rreplace}[r](e_1; e_2) \mapsto \mathsf{rreplace}[r](e_1'; e_2)$ when $e_1 \mapsto e_1'$. The only rule that applies is S-T-Replace, so $\emptyset \vdash e : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$ where:

$$\emptyset \vdash e_1 : \mathsf{stringin}[r_1]$$
$$\emptyset \vdash e_2 : \mathsf{stringin}[r_2]$$

By induction, $\emptyset \vdash e_1' : \mathsf{stringin}[r_1]$. Therefore, $\emptyset \vdash \mathsf{rreplace}[r](r_1'; r_2) : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$ by S-T-Replace.

Case S-E-Replace.

Suppose $e = \mathsf{rreplace}[r](\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[\mathsf{replace}(r; s_1; s_2)]$. The only applicable rule is S-T-Replace, so

$$\emptyset \vdash \mathsf{rstr}[s_1] : \mathsf{stringin}[r_1]$$
$$\emptyset \vdash \mathsf{rstr}[s_2] : \mathsf{stringin}[r_2]$$

By conanical forms, $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$. Therefore,

$$\mathsf{replace}(r; s_1; s_2) \in \mathcal{L}\{\mathsf{lreplace}(r, r_1, r_2)\}$$

by Condition E. It is finally derivable by S-T-Rstr that:

$\emptyset \vdash \mathsf{rstr}[\mathsf{replace}(r; s_1; s_2)] : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]$.

Case S-E-SafeCoerce. Suppose that $\mathsf{rcoerce}[r](\mathsf{rstr}[s_1]) \mapsto \mathsf{rstr}[s_1]$. The only applicable rule is S-T-SafeCoerce, so $\emptyset \vdash \mathsf{rcoerce}[r](s_1) : \mathsf{stringin}[r]$ and $\emptyset \vdash \mathsf{rstr}[s_1] : \mathsf{stringin}[r']$ and $\mathcal{L}\{r'\} \subset \mathcal{L}\{r\}$. By Canonical Forms, $s' \in \mathcal{L}\{r'\}$. By the definition of subset, $s' \in \mathcal{L}\{r\}$. Therefore, by S-T-Rstr, we have that $\emptyset \vdash \mathsf{rstr}[s'] : \mathsf{stringin}[r]$.

Case S-E-Check-Ok. Suppose $\mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) \mapsto [\mathsf{rstr}[s]/x]e_1$ and $s \in \mathcal{L}\{r\}$, and $\emptyset \vdash \mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) : \sigma$. The only rule that applies is S-T-Check, so $\emptyset, x : \mathsf{stringin}[r] \vdash e_1 : \sigma$. By S-T-Rstr, we have that $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]$. By Substitution, we have that $\emptyset \vdash [\mathsf{rstr}[s]/x]e_1 : \sigma$.

Case S-E-Check-NotOk. Suppose $\mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) \mapsto e_2$ and $s \notin \mathcal{L}\{r\}$ and $\emptyset \vdash \mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) : \sigma$. The only applicable rule is S-T-Check, so $\emptyset \vdash e_2 : \sigma$.

□

Theorem J (Type Safety for small step semantics.)**.** *If $\emptyset \vdash e : \sigma$ then either e val or $e \mapsto^* e'$ and $\emptyset \vdash e' : \sigma$.*

Proof. Follows from applying progress and preservation transitively over the multistep judgement. □

3.2.3 The Security Theorem

Theorem 4 (Correctness of Input Sanitation for λ_{RS})**.** *If $\emptyset \vdash e : \mathsf{stringin}[r]$ and $e \mapsto^* \mathsf{rstr}[s]$ then $s \in \mathcal{L}\{r\}$.*

Proof. By type safety, $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]$. By canonical forms, $s \in \mathcal{L}\{r\}$. □

4 λ_P

We will define a translation to a language with only standard strings and regular expressions. The syntax of λ_P is shown in Figure 3.

4.1 Static Semantics

The static semantics of λ_P is shown in Figure 6. The typing context of λ_P obeys the standard structural properties of weakening, exchange and contraction.

4.2 Dynamic Semantics

The dynamic semantics of λ_P is shown in Figure 7.

4.2.1 Canonical Forms

Lemma 5 (Canonical Forms). *If $\emptyset \vdash \dot{v} : \tau$ then:*

- *If $\tau = \tau_1 \to \tau_2$ then $\dot{v} = \lambda x : \tau.\iota$.*
- *If $\tau = \mathsf{regex}$ then $\dot{v} = \mathsf{rx}[r]$.*
- *If $\tau = \mathsf{string}$ then $= \mathsf{str}[s]$.*

Proof. By inspection of the static and dynamic semantics. □

4.2.2 Type Safety

Theorem 6 (Progress). *If $\emptyset \vdash \iota : \tau$ either $\iota = \dot{v}$ or $\iota \mapsto \iota'$.*

Proof. The proof proceeds by induction on the typing assumption.

λ **fragment**. Cases P-T-Var, P-T-Abs, and P-T-App are exactly as in a proof of progress for the simply typed lambda calculus.

P-T-String. In this case, $\iota = \mathsf{str}[s]$, which is a value.

P-T-Regex. In this case, $\iota = \mathsf{rx}[r]$, which is a value.

P-T-Concat. In this case, we have that $\emptyset \vdash \mathsf{concat}(\iota_1; \iota_2) : \mathsf{string}$ and $\emptyset \vdash \iota_1 : \mathsf{string}$ and $\emptyset \vdash \iota_2 : \mathsf{string}$. By the IH, we have that either $\iota_1 \rightsquigarrow \iota_1'$ or $\iota_1 = \dot{v}_1$, and similarly $\iota_2 \rightsquigarrow \iota_2'$ or $\iota_2 = \dot{v}_2$. If ι_1 steps, then we can make progress via PS-E-ConcatLeft. If ι_2 steps, then we can make progress via PS-E-ConcatRight. If both are values, then by canonical forms $\iota_1 = \mathsf{str}[s_1]$ and $\iota_2 = \mathsf{str}[s_2]$ so we can make progress by PS-E-Concat.

P-T-Case. Suppose $\emptyset \vdash \mathsf{strcase}(\iota_1; \iota_2; x, y.\iota_3) : \tau$ and $\emptyset \vdash \iota_1 : \mathsf{string}$. By induction and canonical forms, either $\iota_1 \mapsto \iota_1'$ or $\iota_1 = \mathsf{str}[s_1]$. If ι_1 steps then we can make progress by PS-E-CaseLeft. If it is a value, then by the definition of strings, either $s_1 = \epsilon$ or $s_1 = as$ for some string s. If s_1 is empty, then we can make progress by PS-E-Case-Epsilon. Otherwise, we can make progress by PS-E-Case-Cons.

P-T-Replace. Suppose $\emptyset \vdash \mathsf{replace}(\iota_1; \iota_2; \iota_3) : \mathsf{string}$ and $\emptyset \vdash \iota_1 : \mathsf{regex}$ and $\emptyset \vdash \iota_2 : \mathsf{string}$ and $\emptyset \vdash \iota_3 : \mathsf{string}$. By induction and canonical forms, either $\iota_1 \mapsto \iota_1'$ or $\iota_1 = \mathsf{rx}[r]$. Similarly, $\iota_2 \mapsto \iota_2'$ or $\iota_2 = \mathsf{str}[s_2]$, and $\iota_3 \mapsto \iota_3'$ or $\iota_3 = s_3$. If ι_1 steps, then we can make progress by PS-E-ReplaceLeft. If ι_2 steps then we can make progress by PS-E-ReplaceMid. If ι_3 steps, then we can make progress by PS-E-ReplaceRight. If all three are values, we can make progress by PS-E-Replace.

P-T-Check. Suppose $\emptyset \vdash \mathsf{check}(\iota_1; \iota_2; \iota_3; \iota_4)$ and $\emptyset \vdash \iota_1 : \mathsf{regex}$ and $\emptyset \vdash \iota_2 : \mathsf{string}$. By induction and canonical forms, either $\iota_1 \mapsto \iota_1'$ or $\iota_1 = \mathsf{rx}[r]$. Similarly, $\iota_2 \mapsto \iota_2'$ or $\iota_2 = \mathsf{str}[s]$. If ι_1 steps, then we can make progress by PS-E-CheckLeft. If ι_2 steps, then we can make progress by PS-E-CheckRight. If both are values, then by Assumption B.2, either $s \in \mathcal{L}\{r\}$ or $s \notin \mathcal{L}\{r\}$. In the former case, we can make progress by PS-E-Check-OK. In the latter case, we can make progress by PS-E-Check-NotOK.

□

Assumption K (Substitution)**.** *If* $\Theta, x : \tau' \vdash \iota : \tau$ *and* $\Theta \vdash \iota' : \tau'$ *then* $\Theta \vdash [\iota'/x]\iota : \tau$.

Theorem 7 (Preservation)**.** *If* $\emptyset \vdash \iota : \tau$ *and* $\iota \mapsto \iota'$ *then* $\emptyset \vdash \iota' : \tau$.

Proof. The proof proceeds by induction of the derivations of $\emptyset \vdash \iota : \tau$ and $\iota \mapsto \iota'$.
We treat only the non-lambda fragment.

Case PS-E-ConcatLeft. Suppose:

$$\begin{aligned}
&\iota = \mathsf{rconcat}(\iota_1; \iota_2) \mapsto \mathsf{rconcat}(\iota_1'; \iota_2)\\
&\emptyset \vdash \iota : \mathsf{string}\\
&\iota \mapsto \iota'
\end{aligned}$$

The only applicable typing rule is P-T-Concat, so $\emptyset \vdash \iota_1 : \mathsf{string}$ and $\emptyset \vdash \iota_2 : \mathsf{string}$. By induction, $\emptyset \vdash \iota_1' : \mathsf{string}$, so $\emptyset \vdash \mathsf{rconcat}(\iota_1'; \iota_2) : \mathsf{string}$.

Case PS-E-ConcatRight

$$\begin{aligned}
&e = \mathsf{rconcat}(e_1; e_2) \mapsto \mathsf{rconcat}(e_1; e_2')\\
&\emptyset \vdash e : \mathsf{string}\\
&\iota \mapsto \iota'
\end{aligned}$$

The only applicable typing rule is P-T-Concat, so $\emptyset \vdash \iota_1 : \mathsf{string}$ and $\emptyset \vdash \iota_2 : \mathsf{string}$. By induction, $\emptyset \vdash \iota_1' : \mathsf{string}$, so $\emptyset \vdash \mathsf{rconcat}(\iota_1; \iota_2') : \mathsf{string}$.

Case PS-E-Concat Let $e = \mathsf{rconcat}(\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[s_1 s_2]$. The only rule that applies is P-T-Concat, so $\emptyset \vdash e : \mathsf{string}$. By canonical forms, $\emptyset \mathsf{rstr}[s_1 s_2] : \mathsf{string}$.

Case PS-E-CaseLeft Let $\iota = \mathsf{rstrcase}(\iota_1; \iota_2; x, y.\iota_3) \mapsto \mathsf{rstrcase}(\iota_1'; \iota_2; x, y.\iota_3)$ when $\iota_1 \mapsto \iota_1'$. The only rule that applies is P-T-Case, so $\emptyset \vdash \iota : \tau$ where:

$$\begin{aligned}
\emptyset \vdash \iota_1 &: \mathsf{string}\\
\emptyset \vdash \iota_2 &: \tau\\
\emptyset, x : \mathsf{string}, y : \mathsf{string} \vdash \iota_3 &: \tau
\end{aligned}$$

By induction, $\emptyset \vdash \iota_1' : \mathsf{string}$. By P-T-Case, $\emptyset \vdash \mathsf{rstrcase}(\iota_1; \iota_2; x, y.\iota_3) : \tau$.

Case PS-E-CaseEpsilon Let $\iota = \mathsf{rstrcase}(\mathsf{rstr}[\epsilon]; \iota_2; x, y.\iota_3) \mapsto \iota_2$. The only rule that applies is P-T-Case, so $\emptyset \vdash \iota : \tau$ where $\iota_2 : \tau$.

Case PS-E-Case Let $\iota = \mathsf{rstrcase}(\mathsf{rstr}[as]; \iota_2; x, y.\iota_3) \mapsto [\mathsf{rstr}[a], \mathsf{rstr}[s]/x, y]\iota_3$ The only rule that applies is P-T-Case, so $\emptyset \vdash \iota : \tau$ where:

$$\begin{aligned} \emptyset \vdash \iota_1 &: \mathsf{string} \\ \emptyset \vdash \iota_2 &: \tau \\ \emptyset, x : \mathsf{string}, y : \mathsf{string} \vdash \iota_3 &: \tau \end{aligned}$$

The result follows by the substitution lemma.

Case PS-E-ReplaceLeft Let $\iota = \mathsf{rreplace}[\iota_1](\iota_2; \iota_3) \mapsto \mathsf{rreplace}[\iota_1'](\iota_2; \iota_3)$ where $\iota_1 \mapsto \iota_1'$. The applicable typing rule is P-T-Replace, so $\emptyset \vdash \iota : \mathsf{string}$ where:

$$\begin{aligned} \emptyset \vdash \iota_1 &: \mathsf{regex} \\ \emptyset \vdash \iota_2 &: \mathsf{string} \\ \emptyset \vdash \iota_3 &: \mathsf{string} \end{aligned}$$

By induction, $\emptyset \vdash \iota_1' : \mathsf{regex}$. Therefore, $\emptyset \vdash \mathsf{rreplace}[\iota_1'](\iota_2; \iota_3)$.

Case PS-E-ReplaceMid Let $\iota = \mathsf{rreplace}[\iota_1](\iota_2; \iota_3) \mapsto \mathsf{rreplace}[\iota_1](\iota_2'; \iota_3)$ where $\iota_2 \mapsto \iota_2'$. The applicable typing rule is P-T-Replace, so $\emptyset \vdash \iota : \mathsf{string}$ where:

$$\begin{aligned} \emptyset \vdash \iota_1 &: \mathsf{regex} \\ \emptyset \vdash \iota_2 &: \mathsf{string} \\ \emptyset \vdash \iota_3 &: \mathsf{string} \end{aligned}$$

By induction, $\emptyset \vdash \iota_2' : \mathsf{string}$. Therefore, $\emptyset \vdash \mathsf{rreplace}[\iota_1](\iota_2'; \iota_3)$.

Case PS-E-ReplaceRight Let $\iota = \mathsf{rreplace}[\iota_1](\iota_2; \iota_3) \mapsto \mathsf{rreplace}[\iota_1](\iota_2; \iota_3')$ where $\iota_3 \mapsto \iota_3'$. The applicable typing rule is P-T-Replace, so $\emptyset \vdash \iota : \mathsf{string}$ where:

$$\begin{aligned} \emptyset \vdash \iota_1 &: \mathsf{regex} \\ \emptyset \vdash \iota_2 &: \mathsf{string} \\ \emptyset \vdash \iota_3 &: \mathsf{string} \end{aligned}$$

By induction, $\emptyset \vdash \iota_3' : \mathsf{string}$. Therefore, $\emptyset \vdash \mathsf{rreplace}[\iota_1](\iota_2; \iota_3')$.

Case PS-E-Replace Let $\iota = \mathsf{rreplace}[\mathsf{rx}[r]](\mathsf{rstr}[s_2]; \mathsf{rstr}[s_3]) \mapsto \mathsf{rstr}[\mathsf{lreplace}(r, s_2, s_3)]$. The applicable typing rule is P-T-Replace, so $\emptyset \vdash \iota : \mathsf{string}$. The result follows by canonical forms.

Case PS-E-CheckLeft Let $\iota = \mathsf{rcheck}[\iota_x](\iota_1; \iota_2; \iota_3) \mapsto \mathsf{rcheck}[\iota_x'](\iota_1; \iota_2; \iota_3)$ where $\iota_x \mapsto \iota_x'$. The applicable typing rule is P-T-Check, so $\emptyset \vdash \iota : \tau$ where:

$$\begin{aligned} \emptyset \vdash \iota_x &: \mathsf{regex} \\ \emptyset \vdash \iota_1 &: \mathsf{string} \\ \emptyset \vdash \iota_2 &: \tau \\ \emptyset \vdash \iota_3 &: \tau \end{aligned}$$

By induction, $\iota_x : \mathsf{regex}$. Therefore, $\emptyset \mathsf{rcheck}[\iota_x'](\iota_1; \iota_2; \iota_3) : \tau$.

Case PS-E-CheckRight Let $\iota = \mathsf{rcheck}[\iota_x](\iota_1; \iota_2; \iota_3) \mapsto \mathsf{rcheck}[\iota_x](\iota_1'; \iota_2; \iota_3)$ where $\iota_1 \mapsto \iota_1'$. The applicable typing rule is P-T-Check, so $\emptyset \vdash \iota : \tau$ where:

$$\emptyset \vdash \iota_x : \mathsf{regex}$$
$$\emptyset \vdash \iota_1 : \mathsf{string}$$
$$\emptyset \vdash \iota_2 : \tau$$
$$\emptyset \vdash \iota_3 : \tau$$

By induction, $\iota_1' : \mathsf{string}$. Therefore, $\emptyset \mathsf{rcheck}[\iota_x](\iota_1'; \iota_2; \iota_3) : \tau$.

Case PS-E-Check-Ok Let $\iota = \mathsf{rcheck}[\mathsf{rx}[r]](\mathsf{rstr}[s]; \iota_2; \iota_3) \mapsto \iota_2$ and $s \in \mathcal{L}\{r\}$. The applicable typing rule is P-T-Check, so $\emptyset \vdash \iota : \tau$ where $\emptyset \vdash \iota_2 : \tau$.

Case PS-E-Check-NotOk Let $\iota = \mathsf{rcheck}[\mathsf{rx}[r]](\mathsf{rstr}[s]; \iota_2; \iota_3) \mapsto \iota_3$ where $s \notin \mathcal{L}\{r\}$. The applicable typing rule is P-T-Check, so $\emptyset \vdash \iota : \tau$ where $\emptyset \vdash \iota_3 : \tau$.

□

5 Proofs and Lemmas and Theorems About Translation

Theorem 8 (Translation Correctness)**.** *If $\Psi \vdash e : \sigma$ then there exists an ι such that $[\![e]\!] = \iota$ and $[\![\Psi]\!] \vdash \iota : [\![\sigma]\!]$. Furthermore, if $e \mapsto^* v$ then $\iota \mapsto^* \dot{v}$ such that $[\![v]\!] = \dot{v}$.*

Proof. We present a proof by induction on the structure of e. We write $e \rightsquigarrow \iota$ as shorthand for the final property.

This proof needs to be changed to use only the small-step semantics.

Case $e = \mathsf{rstr}[s]$. Suppose $\Theta \vdash \mathsf{rstr}[s] : \sigma$.

By examination the syntactic structure of conclusions in the relation S-T, we know this is true just in case $\sigma = \mathsf{stringin}[r]$ for some r such that $s \in \mathcal{L}\{r\}$; and of course, there is always such an r.

There are no free variables in $\mathsf{rstr}[s]$, so we might as well proceed from the fact that $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]$.

By definition of the translation ($[\![\cdot]\!]$) the following statements hold:

$$[\![\mathsf{rstr}[s]]\!] = \mathsf{str}[s] \tag{12}$$
$$[\![\mathsf{stringin}[r]]\!] = \mathsf{string} \tag{13}$$
$$[\![\emptyset]\!] = \emptyset \tag{14}$$

Note that $\emptyset \vdash \mathsf{str}[s] : \mathsf{string}$ by P-T-Str. Recall that contexts are standard and, in particular, can be weakened. So since $[\![\Theta]\!]$ is either a weakening of $\emptyset$ or $\emptyset$ itself, $[\![\Theta]\!] \vdash \mathsf{str}[s] : \mathsf{string}$ by weakening.

Summarily, $\mathsf{str}[s]$ is a term of λ_P such that $[\![\Theta]\!] \vdash \mathsf{str}[s] : [\![\sigma]\!]$

It remains to be shown that there exist $v, \dot{v}$ such that $\mathsf{rstr}[s] \mapsto^* v$, $\mathsf{str}[s] \mapsto^* \dot{v}$, and $[\![v]\!] = \dot{v}$. But this is immediate because each term is already a value and $s = s$.

Case $e = \mathsf{rconcat}(e_1; e_2)$. The applicable typing rule is S-T-Concat, so $\Psi \vdash \mathsf{rconcat}(e_1; e_2) : \mathsf{stringin}[r_1 \cdot r_2]$ where $\Psi \vdash e_1 : \mathsf{stringin}[r_1]$ and $\Psi \vdash e_2 : \mathsf{stringin}[r_2]$.

By induction, $e_1 \rightsquigarrow \iota_1$ and $e_2 \rightsquigarrow \iota_2$. Therefore, $[\![\Psi]\!] \vdash \mathsf{concat}(\iota_1; \iota_2)$ by P-T-Concat.

By canonical forms, $e_1 \mapsto^* \mathsf{rstr}[s_1]$ where by induction $\iota_1 \mapsto^* \mathsf{str}[s_1]$. Similarly, $e_2 \mapsto^* \mathsf{rstr}[s_2]$ and $\iota_2 \mapsto^* \mathsf{str}[s_2]$. Therefore, $e \mapsto^* \mathsf{rstr}[s_1 s_2]$ by S-E-Concat at last, and $\mathsf{concat}(\iota_1; \iota_2) \mapsto^* \mathsf{str}[s_1 s_2]$ by P-E-Concat at last. Note that $[\![\mathsf{rstr}[s_1 s_2]]\!] = \mathsf{str}[s_1 s_2]$.

Case $e = \mathsf{rstrcase}(e_1; e_2; x, y.e_3)$. This case relies on our definition of context translation.

Suppose $\Psi \vdash \mathsf{rstrcase}(e_1; e_2; x, y.e_3) : \sigma$. By inversion of the typing relation it follows that $\Psi \vdash e_1 : \mathsf{stringin}[r]$, $\Psi \vdash e_2 : \sigma$ and $\Psi, x : \mathsf{stringin}[\mathsf{lhead}(r)], y : \mathsf{stringin}[\mathsf{ltail}(r)] \vdash e_3 : \sigma$.

By induction, there exists an ι_1 such that $e_1 \mapsto \iota_1$.

By canonical forms, $e_1 \mapsto^* \mathsf{rstr}[s]$. Therefore, $\iota_1 \mapsto *\mathsf{str}[s]$ because $e_1 \rightsquigarrow \iota_1$.

Choose $\iota = \mathsf{strcase}(\iota_1; \iota_2; x, y.\iota_3)$ and note that by the properties established via induction, $[\![e]\!] = \iota$ and $[\![\Psi]\!] \vdash \iota : [\![\sigma]\!]$.

To prove the evaluation correspondence, we consider two cases for the value of s.

Suppose $s = \epsilon$. Then $e \mapsto^* v$ where $e_2 \mapsto^* v$, from which it follows that $\iota \mapsto^* \dot{v}$ where $\iota_2 \mapsto^* \dot{v}$. But recall that $e_2 \rightsquigarrow v_2$ and so $[\![v]\!] = \dot{v}$.

Suppose otherwise that $s = at$ for some character a and string t. Then $e \mapsto^* v$ where $[a, t/x, y]e_3 \mapsto^* v$. Similarly, $\iota \mapsto^* \dot{v}$ where $[a, t/x, y]\iota_3 \mapsto^* \dot{v}$

Case $e = \mathsf{rreplace}[r](e_1; e_2)$. There is only one applicable typing rule, so suppose $\Psi \vdash \mathsf{rreplace}[r](e_1; e_2) : \mathsf{stringin}[\mathsf{lreplace}(r, e_1, e_2)]$. Let $\Theta = [\![\Psi]\!]$. Note that $[\![\mathsf{rreplace}[r](e_1; e_2)]\!] = \mathsf{replace}(\mathsf{rx}[r]; \iota_1; \iota_2)$ where by induction $[\![e_1]\!] = \iota_1$ and $[\![e_2]\!] = \iota_2$ such that $\Theta \vdash \iota_1$ and $\Theta \vdash \iota_2$. It follows by P-T-Replace that $\Theta \vdash \mathsf{replace}(\mathsf{rx}[r]; \iota_1; \iota_2) : \mathsf{string}$. Finally, note that $[\![\mathsf{stringin}[\mathsf{lreplace}(r, e_1, e_2)]]\!] = \mathsf{string}$.

For evaluation correspondence, note that $[\![\mathsf{rstr}[\mathsf{lreplace}(r, s_1, s_2)]]\!] = \mathsf{rstr}[\mathsf{lreplace}(r, s_1, s_2)]$ and so it suffices to show that $\mathsf{replace}(\mathsf{rx}[r]; \iota_1; \iota_2) \mapsto^* \mathsf{rstr}[r] s_1 s_2$. Note that $\mathsf{lreplace}(r, e_1, e_2) \mapsto^* \mathsf{rstr}[\mathsf{lreplace}(r, s_1, s_2)]$ where $e_1 \mapsto^* \mathsf{rstr}[s_1]$, $e_2 \mapsto^* \mathsf{rstr}[s_2]$, $r \mapsto^* r$. By induction, $\iota_1 \mapsto^* \mathsf{rstr}[s_1], \iota_2 \mapsto^* \mathsf{rstr}[s_2]$, and $\mathsf{rx}[r] \mapsto^* \mathsf{rx}[r]$. So by S-E-Replace, the sufficient condition holds.

Case $e = \mathsf{rcoerce}[r](e')$. The only applicable tpying rule is S-T-SafeCoerce, so suppose $\Psi \vdash \mathsf{rcoerce}[r](e') : \mathsf{stringin}[r]$ where $\Psi \vdash e' : \mathsf{stringin}[r']$ and $\mathcal{L}\{r'\} \subseteq \mathcal{L}\{r\}$. By induction, $e' \rightsquigarrow \iota$ for some ι. Therefore, $[\![\mathsf{rcoerce}[r](e')]\!] = \iota$ by Tr-SafeCoerce.

For evaluation correspondence, note that $e \mapsto^* v$ where $e' \mapsto^* v$. The result follows by induction because $e' \rightsquigarrow \iota$.

Case $e = \mathsf{rcheck}[r](e_1; x.e_2; e_3)$. The applicable typing rule is S-T-Check, so $\Psi \vdash e : \sigma$ where $\Psi \vdash e_1 : \mathsf{stringin}[r]$, $\Psi, x : \mathsf{stringin}[r] \vdash e_2 : \sigma$, and $\Psi \vdash e_3 : \sigma$. By induction and a corresponding substitution principle there exists $\iota_1, \iota_2, \iota_3$ such that $e_1 \rightsquigarrow \iota_1$, $e_2 \rightsquigarrow \iota_2$ in context $\Psi, s : \mathsf{stringin}[r]$, and $e_3 \rightsquigarrow \iota_3$. Choose $\iota = \mathsf{check}(\mathsf{rx}[r]; \iota_1; \lambda x.\iota_2; \iota_3)$. The result follows by induction.

□

Theorem 9 (Correctness of Input Sanitation for Translated Terms). *If* $[\![e]\!] = \iota$ *and* $\emptyset \vdash e : \mathsf{stringin}[r]$ *then* $\iota \mapsto^* \mathsf{str}[s]$ *for* $s \in \mathcal{L}\{r\}$.

Proof. By 4, $e \mapsto^* \mathsf{rstr}[s]$ where $\emptyset \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]$. Therefore, $s \in \mathcal{L}\{r\}$. Note that $[\![\cdot]\!]$ is a function and $[\![\mathsf{rstr}[s]]\!] = \mathsf{str}[s]$; therefore, by theorem 8, $\iota \mapsto^* \mathsf{str}[s]$. □

References

[1] N. Fulton, C. Omar, and J. Aldrich. Statically typed string sanitation inside a python. SPLASH '14. ACM, 2014.

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

Figure 1: Syntax of regular expressions over the alphabet Σ.

$$\begin{array}{lll} \sigma & ::= \ \sigma \to \sigma \mid \mathsf{stringin}[r] & \text{source types} \\ e & ::= \ x \mid v & \text{source terms} \\ & \mid \ \mathsf{rconcat}(e; e) \mid \mathsf{rstrcase}(e; e; x, y.e) & s \in \Sigma^* \\ & \mid \ \mathsf{rreplace}[r](e; e) \mid \mathsf{rcoerce}[r](e) \mid \mathsf{rcheck}[r](e; x.e; e) & \\ v & ::= \ \lambda x.e \mid \mathsf{rstr}[s] & \text{source values} \end{array}$$

Figure 2: Syntax of λ_{RS}

$$\begin{array}{lll} \tau & ::= \ \tau \to \tau \mid \mathsf{string} \mid \mathsf{regex} & \text{target types} \\ \iota & ::= \ x \mid \dot{v} & \text{target terms} \\ & \mid \ \mathsf{concat}(\iota; \iota) \mid \mathsf{strcase}(\iota; \iota; x, y.\iota) & \\ & \mid \ \mathsf{rx}[r] \mid \mathsf{replace}(\iota; \iota; \iota) \mid \mathsf{check}(\iota; \iota; \iota; \iota) & \\ \dot{v} & ::= \ \lambda x.\iota \mid \mathsf{str}[s] \mid \mathsf{rx}[r] & \text{target values} \end{array}$$

Figure 3: Syntax of λ_P

$$\boxed{\Psi \vdash e : \sigma} \qquad \Psi ::= \emptyset \mid \Psi, x : \sigma$$

$$\textsc{S-T-Var}\ \frac{x : \sigma \in \Psi}{\Psi \vdash x : \sigma} \qquad \textsc{S-T-Abs}\ \frac{\Psi, x : \sigma_1 \vdash e : \sigma_2}{\Psi \vdash \lambda x.e : \sigma_1 \to \sigma_2} \qquad \textsc{S-T-App}\ \frac{\Psi \vdash e_1 : \sigma_2 \to \sigma \qquad \Psi \vdash e_2 : \sigma_2}{\Psi \vdash e_1(e_2) : \sigma} \qquad \textsc{S-T-Stringin-I}\ \frac{s \in \mathcal{L}\{r\}}{\Psi \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]}$$

$$\textsc{S-T-Concat}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r_1] \qquad \Psi \vdash e_2 : \mathsf{stringin}[r_2]}{\Psi \vdash \mathsf{rconcat}(e_1; e_2) : \mathsf{stringin}[r_1 \cdot r_2]}$$

$$\textsc{S-T-Case}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r] \qquad \Psi \vdash e_2 : \sigma \qquad \Psi, x : \mathsf{stringin}[\mathsf{lhead}(r)], y : \mathsf{stringin}[\mathsf{ltail}(r)] \vdash e_3 : \sigma}{\Psi \vdash \mathsf{rstrcase}(e_1; e_2; x, y.e_3) : \sigma}$$

$$\textsc{S-T-Replace}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r_1] \qquad \Psi \vdash e_2 : \mathsf{stringin}[r_2]}{\Psi \vdash \mathsf{rreplace}[r](e_1; e_2) : \mathsf{stringin}[\mathsf{lreplace}(r, r_1, r_2)]} \qquad \textsc{S-T-SafeCoerce}\ \frac{\Psi \vdash e : \mathsf{stringin}[r'] \qquad \mathcal{L}\{r'\} \subseteq \mathcal{L}\{r\}}{\Psi \vdash \mathsf{rcoerce}[r](e) : \mathsf{stringin}[r]}$$

$$\textsc{S-T-Check}\ \frac{\Psi \vdash e_0 : \mathsf{stringin}[r] \qquad \Psi, x : \mathsf{stringin}[r] \vdash e_1 : \sigma \qquad \Psi \vdash e_2 : \sigma}{\Psi \vdash \mathsf{rcheck}[r](e_0; x.e_1; e_2) : \sigma}$$

Figure 4: Typing rules for λ_{RS}. The typing context Ψ is standard.

$\boxed{e \mapsto e}$

$$\frac{e_1 \mapsto e_1'}{e_1(e_2) \mapsto e_1'(e_2)}\ \text{SS-E-AppLeft} \qquad \frac{e_2 \mapsto e_2'}{v_1(e_2) \mapsto v_1(e_2')}\ \text{SS-E-AppRight} \qquad \frac{}{(\lambda x : \sigma.e)v_2 \mapsto [v_2/x]e}\ \text{SS-E-AppAbs}$$

$$\frac{e_1 \mapsto e_1'}{\mathsf{rconcat}(e_1; e_2) \mapsto \mathsf{rconcat}(e_1'; e_2)}\ \text{SS-E-Concat-Left}$$

$$\frac{e_2 \mapsto e_2'}{\mathsf{rconcat}(v_1; e_2) \mapsto \mathsf{rconcat}(v_1; e_2')}\ \text{SS-E-Concat-Right} \qquad \frac{}{\mathsf{rconcat}(\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[s_1 s_2]}\ \text{SS-E-Concat}$$

$$\frac{e_1 \mapsto e_1'}{\mathsf{rstrcase}(e_1; e_2; x, y.e_3) \mapsto \mathsf{rstrcase}(e_1'; e_2; x, y.e_3)}\ \text{SS-E-Case-Left} \qquad \frac{}{\mathsf{rstrcase}(\mathsf{rstr}[\epsilon]; e_2; x.y.e_3) \mapsto e_2}\ \text{SS-E-Case-}\epsilon\text{-val}$$

$$\frac{}{\mathsf{rstrcase}(\mathsf{rstr}[as]; e_2; x, y.e_3) \mapsto [\mathsf{rstr}[a], \mathsf{rstr}[s]/x, y]e_3}\ \text{SS-E-Case-Concat} \qquad \frac{e_1 \mapsto e_1'}{\mathsf{rreplace}[r](v_1; e_2) \mapsto \mathsf{rreplace}[r](v_1'; e_2)}\ \text{SS-E-Replace-Left}$$

$$\frac{e_2 \mapsto e_2'}{\mathsf{rreplace}[r](e_1; e_2) \mapsto \mathsf{rreplace}[r](e_1; e_2')}\ \text{SS-E-Replace-Right} \qquad \frac{}{\mathsf{rreplace}[r](\mathsf{rstr}[s_1]; \mathsf{rstr}[s_2]) \mapsto \mathsf{rstr}[\mathsf{replace}(r; s_1; s_2)]}\ \text{SS-E-Replace}$$

$$\frac{e \mapsto e'}{\mathsf{rcoerce}[r](e) \mapsto \mathsf{rcoerce}[r](e')}\ \text{SS-E-SafeCoerce-Step} \qquad \frac{}{\mathsf{rcoerce}[r](\mathsf{rstr}[s]) \mapsto \mathsf{rstr}[s]}\ \text{SS-E-SafeCoerce}$$

$$\frac{e \mapsto e'}{\mathsf{rcheck}[r](e; x.e_1; e_2) \mapsto \mathsf{rcheck}[r](e'; x.e_1; e_2)}\ \text{SS-E-Check-StepLeft}$$

$$\frac{s \in \mathcal{L}\{r\}}{\mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) \mapsto [\mathsf{rstr}[s]/x]e_1}\ \text{SS-E-Check-Ok} \qquad \frac{s \notin \mathcal{L}\{r\}}{\mathsf{rcheck}[r](\mathsf{rstr}[s]; x.e_1; e_2) \mapsto e_2}\ \text{SS-E-Check-NotOk}$$

Figure 5: Small step semantics for λ_{RS}.

$\boxed{\Theta \vdash \iota : \tau}$ $\quad \Theta ::= \emptyset \mid \Theta, x : \tau$

$$\textsc{P-T-Var}\ \frac{x : \tau \in \Theta}{\Theta \vdash x : \tau} \qquad \textsc{P-T-Abs}\ \frac{\Theta, x : \tau_1 \vdash \iota_2 : \tau_2}{\Theta \vdash \lambda x.\iota_2 : \tau_1 \to \tau_2} \qquad \textsc{P-T-App}\ \frac{\Theta \vdash \iota_1 : \tau_2 \to \tau \qquad \Theta \vdash \iota_2 : \tau_2}{\Theta \vdash \iota_1(\iota_2) : \iota} \qquad \textsc{P-T-String}\ \frac{}{\Theta \vdash \mathsf{str}[s] : \mathsf{string}}$$

$$\textsc{P-T-Regex}\ \frac{}{\Theta \vdash \mathsf{rx}[r] : \mathsf{regex}} \qquad \textsc{P-T-Concat}\ \frac{\Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta \vdash \iota_2 : \mathsf{string}}{\Theta \vdash \mathsf{concat}(\iota_1; \iota_2) : \mathsf{string}}$$

$$\textsc{P-T-Case}\ \frac{\Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta \vdash \iota_2 : \tau \qquad \Theta, x : \mathsf{string}, y : \mathsf{string} \vdash \iota_3 : \tau}{\Theta \vdash \mathsf{strcase}(\iota_1; \iota_2; x, y.\iota_3) : \tau}$$

$$\textsc{P-T-Replace}\ \frac{\Theta \vdash \iota_1 : \mathsf{regex} \qquad \Theta \vdash \iota_2 : \mathsf{string} \qquad \Theta \vdash \iota_3 : \mathsf{string}}{\Theta \vdash \mathsf{replace}(\iota_1; \iota_2; \iota_3) : \mathsf{string}}$$

$$\textsc{P-T-Check}\ \frac{\Theta \vdash \iota_x : \mathsf{regex} \qquad \Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta \vdash \iota_2 : \tau \qquad \Theta \vdash \iota_3 : \tau}{\Theta \vdash \mathsf{check}(\iota_x; \iota_1; \iota_2; \iota_3) : \tau}$$

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

$\boxed{\iota \mapsto \iota}$

$$\text{PS-E-AppLeft}\quad \frac{\iota_1 \mapsto \iota_1'}{\iota_1(\iota_2) \mapsto \iota_1'(\iota_2)}$$

$$\text{PS-E-AppRight}\quad \frac{\iota_2 \mapsto \iota_2'}{\dot{v}_1(\iota_2) \mapsto \dot{v}_1(\iota_2')}$$

$$\text{PS-E-AppAbs}\quad \frac{}{(\lambda x : \tau.\iota)\dot{v}_2 \mapsto [\dot{v}_2/x]\iota}$$

$$\text{PS-E-ConcatLeft}\quad \frac{\iota_1 \mapsto \iota_1'}{\mathsf{concat}(\iota_1; \iota_2) \mapsto \mathsf{concat}(\iota_1'; \iota_2)}$$

$$\text{PS-E-ConcatRight}\quad \frac{\iota_2 \mapsto \iota_2'}{\mathsf{concat}(\mathsf{str}[s_1]; \iota_2) \mapsto \mathsf{concat}(\mathsf{str}[s_1]; \iota_2')}$$

$$\text{PS-E-Concat}\quad \frac{}{\mathsf{concat}(\mathsf{str}[s_1]; \mathsf{str}[s_2]) \mapsto \mathsf{str}[s_1 s_2]}$$

$$\text{PS-E-CaseLeft}\quad \frac{\iota_1 \mapsto \iota_1'}{\mathsf{strcase}(\iota_1; \iota_2; x, y.\iota_3) \mapsto \mathsf{strcase}(\iota_1'; \iota_2; x, y.\iota_3)}$$

$$\text{PS-E-Case-Epsilon}\quad \frac{}{\mathsf{strcase}(\mathsf{str}[\epsilon]; \iota_2; x, y.\iota_3) \mapsto \iota_2}$$

$$\text{PS-E-Case}\quad \frac{}{\mathsf{strcase}(\mathsf{str}[as]; \iota_2; x, y.\iota_3) \mapsto [\mathsf{str}[a], \mathsf{str}[s]/x, y]\iota_3}$$

$$\text{PS-E-ReplaceLeft}\quad \frac{\iota_1 \mapsto \iota_1'}{\mathsf{replace}(\iota_1; \iota_2; \iota_3) \mapsto \mathsf{replace}(\iota_1'; \iota_2; \iota_3)}$$

$$\text{PS-E-ReplaceMid}\quad \frac{\iota_2 \mapsto \iota_2'}{\mathsf{replace}(\mathsf{rx}[r]; \iota_2; \iota_3) \mapsto \mathsf{replace}(\mathsf{rx}[r]; \iota_2'; \iota_3)}$$

$$\text{PS-E-ReplaceRight}\quad \frac{\iota_3 \mapsto \iota_3'}{\mathsf{replace}(\mathsf{rx}[r]; \mathsf{str}[s_2]; \iota_3) \mapsto \mathsf{replace}(\mathsf{rx}[r]; \mathsf{str}[s_2]; \iota_3')}$$

$$\text{PS-E-Replace}\quad \frac{}{\mathsf{replace}(\mathsf{rx}[r]; \mathsf{str}[s_2]; \mathsf{str}[s_3]) \mapsto \mathsf{str}[\mathsf{replace}(r; s_2; s_3)]}$$

$$\text{PS-E-CheckLeft}\quad \frac{\iota_1 \mapsto \iota_1'}{\mathsf{check}(\iota_1; \iota_2; \iota_3; \iota_4) \mapsto \mathsf{check}(\iota_1'; \iota_2; \iota_3; \iota_4)}$$

$$\text{PS-E-CheckRight}\quad \frac{\iota_2 \mapsto \iota_2'}{\mathsf{check}(\mathsf{rx}[r]; \iota_2; \iota_3; \iota_4) \mapsto \mathsf{check}(\mathsf{rx}[r]; \iota_2'; \iota_3; \iota_4)}$$

$$\text{PS-E-Check-OK}\quad \frac{s \in \mathcal{L}\{r\}}{\mathsf{check}(\mathsf{rx}[r]; \mathsf{str}[s]; \iota_3; \iota_4) \mapsto \iota_3}$$

$$\text{PS-E-Check-NotOK}\quad \frac{s \notin \mathcal{L}\{r\}}{\mathsf{check}(\mathsf{rx}[r]; \mathsf{str}[s]; \iota_3; \iota_4) \mapsto \iota_4}$$

Figure 7: Small step semantics for λ_P

.

$\boxed{\llbracket \sigma \rrbracket = \tau}$

$$\textsc{Tr-T-String}\quad \frac{}{\llbracket \mathsf{stringin}[r] \rrbracket = \mathsf{string}} \qquad \textsc{Tr-T-Arrow}\quad \frac{\llbracket \sigma_1 \rrbracket = \tau_1 \qquad \llbracket \sigma_2 \rrbracket = \tau_2}{\llbracket \sigma_1 \rightarrow \sigma_2 \rrbracket = \tau_1 \rightarrow \tau_2}$$

$\boxed{\llbracket \Psi \rrbracket = \Theta}$

$$\textsc{Tr-T-Context-Emp}\quad \frac{}{\llbracket \emptyset \rrbracket = \emptyset} \qquad \textsc{Tr-T-Context-Ext}\quad \frac{\llbracket \Psi \rrbracket = \Theta \qquad \llbracket \sigma \rrbracket = \tau}{\llbracket \Psi, x : \sigma \rrbracket = \Theta, x : \tau}$$

$\boxed{\llbracket e \rrbracket = \iota}$

$$\textsc{Tr-Var}\quad \frac{}{\llbracket x \rrbracket = x} \qquad \textsc{Tr-Abs}\quad \frac{\llbracket e \rrbracket = \iota}{\llbracket \lambda x.e \rrbracket = \lambda x.\iota} \qquad \textsc{Tr-App}\quad \frac{\llbracket e_1 \rrbracket = \iota_1 \qquad \llbracket e_2 \rrbracket = \iota_2}{\llbracket e_1(e_2) \rrbracket = \iota_1(\iota_2)}$$

$$\textsc{Tr-Case}\quad \frac{\llbracket e_1 \rrbracket = \iota_1 \qquad \llbracket e_2 \rrbracket = \iota_2 \qquad \llbracket e_3 \rrbracket = \iota_3}{\llbracket \mathsf{rstrcase}(e_1; e_2; x, y.e_3) \rrbracket = \mathsf{strcase}(\iota_1; \iota_2; x, y.\iota_3)}$$

$$\textsc{Tr-string}\quad \frac{}{\llbracket \mathsf{rstr}[s] \rrbracket = \mathsf{str}[s]} \qquad \textsc{Tr-Concat}\quad \frac{\llbracket e_1 \rrbracket = \iota_1 \qquad \llbracket e_2 \rrbracket = \iota_2}{\llbracket \mathsf{rconcat}(e_1; e_2) \rrbracket = \mathsf{concat}(\iota_1; \iota_2)}$$

$$\textsc{Tr-Subst}\quad \frac{\llbracket e_1 \rrbracket = \iota_1 \qquad \llbracket e_2 \rrbracket = \iota_2}{\llbracket \mathsf{rreplace}[r](e_1; e_2) \rrbracket = \mathsf{replace}(\mathsf{rx}[r]; \iota_1; \iota_2)}$$

$$\textsc{Tr-SafeCoerce}\quad \frac{\llbracket e \rrbracket = \iota}{\llbracket \mathsf{rcoerce}[r'](e) \rrbracket = \iota} \qquad \textsc{Tr-Check}\quad \frac{\llbracket e \rrbracket = \iota \qquad \llbracket e_1 \rrbracket = \iota_1 \qquad \llbracket e_2 \rrbracket = \iota_2}{\llbracket \mathsf{rcheck}[r](e; x.e_1; e_2) \rrbracket = \mathsf{check}(\mathsf{rx}[r]; \iota; (\lambda x.\iota_1)(\iota); \iota_2)}$$

Figure 8: Translation from λ_{RS} to λ_P