POSIX Lexing with Bitcoded Derivatives

Chengsong Tan ✉
King's College London

Christian Urban ✉
King's College London

Abstract

Sulzmann and Lu describe a lexing algorithm that calculates Brzozowski derivatives using bitcodes annotated to regular expressions. Their algorithm generates POSIX values which encode the information of *how* a regular expression matches a string—that is, which part of the string is matched by which part of the regular expression. This information is needed in the context of lexing in order to extract and to classify tokens. The purpose of the bitcodes is to generate POSIX values incrementally while derivatives are calculated. They also help with designing an "aggressive" simplification function that keeps the size of derivatives finite. Without simplification the size of some derivatives can grow arbitrarily big resulting in an extremely slow lexing algorithm. In this paper we describe a variant of Sulzmann and Lu's algorithm: Our variant is a recursive functional program, whereas Sulzmann and Lu's version involves a fixpoint construction. We *(i)* prove in Isabelle/HOL that our algorithm is correct and generates unique POSIX values; we also *(ii)* establish a finite bound for the size of the derivatives.

2012 ACM Subject Classification Design and analysis of algorithms; Formal languages and automata theory

Keywords and phrases POSIX matching and lexing, derivatives of regular expressions, Isabelle/HOL

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

In the last fifteen or so years, Brzozowski's derivatives of regular expressions have sparked quite a bit of interest in the functional programming and theorem prover communities. The beauty of Brzozowski's derivatives [4] is that they are neatly expressible in any functional language, and easily definable and reasoned about in theorem provers—the definitions just consist of inductive datatypes and simple recursive functions. Derivatives of a regular expression, written $r\backslash c$, give a simple solution to the problem of matching a string s with a regular expression r: if the derivative of r w.r.t. (in succession) all the characters of the string matches the empty string, then r matches s (and *vice versa*). We are aware of a mechanised correctness proof of Brzozowski's derivative-based matcher in HOL4 by Owens and Slind [10]. Another one in Isabelle/HOL is part of the work by Krauss and Nipkow [7]. And another one in Coq is given by Coquand and Siles [5]. Also Ribeiro and Du Bois give one in Agda [11].

However, there are two difficulties with derivative-based matchers: First, Brzozowski's original matcher only generates a yes/no answer for whether a regular expression matches a string or not. This is too little information in the context of lexing where separate tokens must be identified and also classified (for example as keywords or identifiers). Sulzmann and Lu [12] overcome this difficulty by cleverly extending Brzozowski's matching algorithm. Their extended version generates additional information on *how* a regular expression matches a string following the POSIX rules for regular expression matching. They achieve this by adding a second "phase" to Brzozowski's algorithm involving an injection function. In our own earlier work we provided the formal specification of what POSIX matching means and proved in Isabelle/HOL the correctness of Sulzmann and Lu's extended algorithm accordingly [3].

The second difficulty is that Brzozowski's derivatives can grow to arbitrarily big sizes. For example if we start with the regular expression $(a + aa)^*$ and take successive derivatives according to the character a, we end up with a sequence of ever-growing derivatives like

Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$$
\begin{array}{lll}
(a+aa)^* & \xrightarrow{\_\backslash a} & (\mathbf{1}+\mathbf{1}a)\cdot(a+aa)^* \\
& \xrightarrow{\_\backslash a} & (\mathbf{0}+\mathbf{0}a+\mathbf{1})\cdot(a+aa)^* \;+\; (\mathbf{1}+\mathbf{1}a)\cdot(a+aa)^* \\
& \xrightarrow{\_\backslash a} & (\mathbf{0}+\mathbf{0}a+\mathbf{0})\cdot(a+aa)^* + (\mathbf{1}+\mathbf{1}a)\cdot(a+aa)^* + \\
& & \quad (\mathbf{0}+\mathbf{0}a+\mathbf{1})\cdot(a+aa)^* + (\mathbf{1}+\mathbf{1}a)\cdot(a+aa)^* \\
& \xrightarrow{\_\backslash a} & \ldots \qquad \text{(regular expressions of sizes 98, 169, 283, 468, 767, \ldots)}
\end{array}
$$

where after around 35 steps we run out of memory on a typical computer (we shall define shortly the precise details of our regular expressions and the derivative operation). Clearly, the notation involving $\mathbf{0}$s and $\mathbf{1}$s already suggests simplification rules that can be applied to regular regular expressions, for example $\mathbf{0}\,r \Rightarrow \mathbf{0}$, $\mathbf{1}\,r \Rightarrow r$, $\mathbf{0} + r \Rightarrow r$ and $r + r \Rightarrow r$. While such simple-minded simplifications have been proved in our earlier work to preserve the correctness of Sulzmann and Lu's algorithm [3], they unfortunately do *not* help with limiting the growth of the derivatives shown above: the growth is slowed, but the derivatives can still grow rather quickly beyond any finite bound.

Sulzmann and Lu overcome this "growth problem" in a second algorithm [12] where they introduce bitcoded regular expressions. In this version, POSIX values are represented as bitsequences and such sequences are incrementally generated when derivatives are calculated. The compact representation of bitsequences and regular expressions allows them to define a more "aggressive" simplification method that keeps the size of the derivatives finite no matter what the length of the string is. They make some informal claims about the correctness and linear behaviour of this version, but do not provide any supporting proof arguments, not even "pencil-and-paper" arguments. They write about their bitcoded *incremental parsing method* (that is the algorithm to be formalised in this paper):

> *"Correctness Claim: We further claim that the incremental parsing method [..] in combination with the simplification steps [..] yields POSIX parse trees. We have tested this claim extensively [..] but yet have to work out all proof details." [12, Page 14]*

Contributions: We have implemented in Isabelle/HOL the derivative-based lexing algorithm of Sulzmann and Lu [12] where regular expressions are annotated with bitsequences. We define the crucial simplification function as a recursive function, without the need of a fix-point operation. One objective of the simplification function is to remove duplicates of regular expressions. For this Sulzmann and Lu use in their paper the standard *nub* function from Haskell's list library, but this function does not achieve the intended objective with bitcoded regular expressions. The reason is that in the bitcoded setting, each copy generally has a different bitcode annotation—so *nub* would never "fire". Inspired by Scala's library for lists, we shall instead use a *distinctBy* function that finds duplicates under an erasing function which deletes bitcodes. We shall also introduce our own argument and definitions for establishing the correctness of the bitcoded algorithm when simplifications are included.

In this paper, we shall first briefly introduce the basic notions of regular expressions and describe the basic definitions of POSIX lexing from our earlier work [3]. This serves as a reference point for what correctness means in our Isabelle/HOL proofs. We shall then prove the correctness for the bitcoded algorithm without simplification, and after that extend the proof to include simplification.

2 Background

In our Isabelle/HOL formalisation strings are lists of characters with the empty string being represented by the empty list, written [], and list-cons being written as $\_ :: \_$; string concatenation is $\_ @ \_$. We often use the usual bracket notation for lists also for strings; for example a string consisting of just a single character c is written $[c]$. Our regular expressions are defined as usual as the elements of the following inductive datatype:

$$r ::= \ \mathbf{0} \mid \mathbf{1} \mid c \mid r_1 + r_2 \mid r_1 \cdot r_2 \mid r^*$$

where $\mathbf{0}$ stands for the regular expression that does not match any string, $\mathbf{1}$ for the regular expression that matches only the empty string and c for matching a character literal. The constructors $+$ and $\cdot$ represent alternatives and sequences, respectively. The *language* of a regular expression, written L, is defined as usual and we omit giving the definition here (see for example [3]).

Central to Brzozowski's regular expression matcher are two functions called *nullable* and *derivative*. The latter is written $r\backslash c$ for the derivative of the regular expression r w.r.t. the character c. Both functions are defined by recursion over regular expressions.

$$\begin{array}{rcl}
\mathbf{0}\backslash c & \stackrel{\text{def}}{=} & \mathbf{0}\\
\mathbf{1}\backslash c & \stackrel{\text{def}}{=} & \mathbf{0}\\
d\backslash c & \stackrel{\text{def}}{=} & \textit{if } c = d \textit{ then } \mathbf{1} \textit{ else } \mathbf{0}\\
(r_1 + r_2)\backslash c & \stackrel{\text{def}}{=} & (r_1\backslash c) + (r_2\backslash c)\\
(r_1 \cdot r_2)\backslash c & \stackrel{\text{def}}{=} & \textit{if nullable } r_1\\
 & & \textit{then } (r_1\backslash c) \cdot r_2 + (r_2\backslash c)\\
 & & \textit{else } (r_1\backslash c) \cdot r_2\\
(r^*)\backslash c & \stackrel{\text{def}}{=} & (r\backslash c) \cdot r^*
\end{array}$$

$$\begin{array}{lcl}
\textit{nullable}\ (\mathbf{0}) & \stackrel{\text{def}}{=} & \textit{False}\\
\textit{nullable}\ (\mathbf{1}) & \stackrel{\text{def}}{=} & \textit{True}\\
\textit{nullable}\ (c) & \stackrel{\text{def}}{=} & \textit{False}\\
\textit{nullable}\ (r_1 + r_2) & \stackrel{\text{def}}{=} & \textit{nullable}\ r_1 \vee \textit{nullable}\ r_2\\
\textit{nullable}\ (r_1 \cdot r_2) & \stackrel{\text{def}}{=} & \textit{nullable}\ r_1 \wedge \textit{nullable}\ r_2\\
\textit{nullable}\ (r^*) & \stackrel{\text{def}}{=} & \textit{True}
\end{array}$$

We can extend this definition to give derivatives w.r.t. strings:

$$r\backslash[] \stackrel{\text{def}}{=} r \qquad r\backslash(c :: s) \stackrel{\text{def}}{=} (r\backslash c)\backslash s$$

Using *nullable* and the derivative operation, we can define the following simple regular expression matcher:

$$\textit{match}\ s\ r \stackrel{\text{def}}{=} \textit{nullable}(r\backslash s)$$

This is essentially Brzozowski's algorithm from 1964. Its main virtue is that the algorithm can be easily implemented as a functional program (either in a functional programming language or in a theorem prover). The correctness proof for *match* amounts to establishing the property

▶ **Proposition 1.** *match* $s\ r$ *if and only if* $s \in L(r)$

It is a fun exercise to formally prove this property in a theorem prover.

The novel idea of Sulzmann and Lu is to extend this algorithm for lexing, where it is important to find out which part of the string is matched by which part of the regular expression. For this Sulzmann and Lu presented two lexing algorithms in their paper [12]. The first algorithm consists of two phases: first a matching phase (which is Brzozowski's algorithm) and then a value construction phase. The values encode *how* a regular expression matches a string. *Values* are defined as the inductive datatype

$$v := \textit{Empty} \mid \textit{Char}\ c \mid \textit{Left}\ v \mid \textit{Right}\ v \mid \textit{Seq}\ v_1\ v_2 \mid \textit{Stars}\ vs$$

where we use *vs* to stand for a list of values. The string underlying a value can be calculated by a *flat* function, written $|\_|$. It traverses a value and collects the characters contained in it. Sulzmann and Lu also define inductively an inhabitation relation that associates values to regular expressions:

$$\frac{}{\vdash \textit{Empty} : \mathbf{1}} \qquad \frac{}{\vdash \textit{Char}\ c : c}$$

$$\frac{\vdash v_1 : r_1}{\vdash \textit{Left}\ v_1 : r_1 + r_2} \qquad \frac{\vdash v_2 : r_2}{\vdash \textit{Right}\ v_2 : r_1 + r_2}$$

$$\frac{\vdash v_1 : r_1 \qquad \vdash v_2 : r_2}{\vdash \textit{Seq}\ v_1\ v_2 : r_1 \cdot r_2} \qquad \frac{\forall v \in vs. \vdash v : r \wedge |v| \neq []}{\vdash \textit{Stars}\ vs : r^*}$$

$$\frac{}{([], \mathbf{1}) \rightarrow \mathit{Empty}}P\mathbf{1} \qquad \frac{}{([c], c) \rightarrow \mathit{Char}\ c}Pc$$

$$\frac{(s, r_1) \rightarrow v}{(s, r_1 + r_2) \rightarrow \mathit{Left}\ v}P{+}L \qquad \frac{(s, r_2) \rightarrow v \qquad s \notin L\ r_1}{(s, r_1 + r_2) \rightarrow \mathit{Right}\ v}P{+}R$$

$$\frac{\begin{array}{c}(s_1, r_1) \rightarrow v_1 \qquad (s_2, r_2) \rightarrow v_2 \\ \nexists s_3\ s_4.\ s_3 \neq [] \wedge s_3 \mathbin{@} s_4 = s_2 \wedge s_1 \mathbin{@} s_3 \in L\ r_1 \wedge s_4 \in L\ r_2\end{array}}{(s_1 \mathbin{@} s_2, r_1 \cdot r_2) \rightarrow \mathit{Seq}\ v_1\ v_2}PS$$

$$\frac{}{([], r^*) \rightarrow \mathit{Stars}\ []}P[] \qquad \frac{\begin{array}{c}(s_1, r) \rightarrow v \qquad (s_2, r^*) \rightarrow \mathit{Stars}\ vs \qquad |v| \neq [] \\ \nexists s_3\ s_4.\ s_3 \neq [] \wedge s_3 \mathbin{@} s_4 = s_2 \wedge s_1 \mathbin{@} s_3 \in L\ r \wedge s_4 \in L\ (r^*)\end{array}}{(s_1 \mathbin{@} s_2, r^*) \rightarrow \mathit{Stars}\ (v :: vs)}P\star$$

Figure 1 The inductive definition of POSIX values taken from our earlier paper [3]. The ternary relation, written $(s, r) \rightarrow v$, formalises the notion of given a string s and a regular expression r what is the unique value v that satisfies the informal POSIX constraints for regular expression matching.

Note that no values are associated with the regular expression $\mathbf{0}$, since it cannot match any string. It is routine to establish how values "inhabiting" a regular expression correspond to the language of a regular expression, namely

▶ **Proposition 2.** $L\ r = \{|v| \mid \vdash v : r\}$

In general there is more than one value inhabited by a regular expression (meaning regular expressions can typically match more than one string). But even when fixing a string from the language of the regular expression, there are generally more than one way of how the regular expression can match this string. POSIX lexing is about identifying the unique value for a given regular expression and a string that satisfies the informal POSIX rules (see [1, 8, 9, 12, 13]).[1] Sometimes these informal rules are called *maximal much rule* and *rule priority*. One contribution of our earlier paper is to give a convenient specification for what POSIX values are (the inductive rules are shown in Figure 1).

The clever idea by Sulzmann and Lu [12] in their first algorithm is to define an injection function on values that mirrors (but inverts) the construction of the derivative on regular expressions. Essentially it injects back a character into a value. For this they define two functions called *mkeps* and *inj*:

$$\begin{array}{lcl}
\mathit{mkeps}\ \mathbf{1} & \stackrel{\text{def}}{=} & \mathit{Empty} \\
\mathit{mkeps}\ (r_1 \cdot r_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\ (\mathit{mkeps}\ r_1)\ (\mathit{mkeps}\ r_2) \\
\mathit{mkeps}\ (r_1 + r_2) & \stackrel{\text{def}}{=} & \mathit{if}\ \mathit{nullable}\ r_1\ \mathit{then}\ \mathit{Left}\ (\mathit{mkeps}\ r_1)\ \mathit{else}\ \mathit{Right}\ (\mathit{mkeps}\ r_2) \\
\mathit{mkeps}\ (r^*) & \stackrel{\text{def}}{=} & \mathit{Stars}\ []
\end{array}$$

$$\begin{array}{lcl}
\mathit{inj}\ d\ c\ (\mathit{Empty}) & \stackrel{\text{def}}{=} & \mathit{Char}\ d \\
\mathit{inj}\ (r_1 + r_2)\ c\ (\mathit{Left}\ v_1) & \stackrel{\text{def}}{=} & \mathit{Left}\ (\mathit{inj}\ r_1\ c\ v_1) \\
\mathit{inj}\ (r_1 + r_2)\ c\ (\mathit{Right}\ v_2) & \stackrel{\text{def}}{=} & \mathit{Right}\ (\mathit{inj}\ r_2\ c\ v_2) \\
\mathit{inj}\ (r_1 \cdot r_2)\ c\ (\mathit{Seq}\ v_1\ v_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\ (\mathit{inj}\ r_1\ c\ v_1)\ v_2 \\
\mathit{inj}\ (r_1 \cdot r_2)\ c\ (\mathit{Left}\ (\mathit{Seq}\ v_1\ v_2)) & \stackrel{\text{def}}{=} & \mathit{Seq}\ (\mathit{inj}\ r_1\ c\ v_1)\ v_2 \\
\mathit{inj}\ (r_1 \cdot r_2)\ c\ (\mathit{Right}\ v_2) & \stackrel{\text{def}}{=} & \mathit{Seq}\ (\mathit{mkeps}\ r_1)\ (\mathit{inj}\ r_2\ c\ v_2) \\
\mathit{inj}\ (r^*)\ c\ (\mathit{Seq}\ v\ (\mathit{Stars}\ vs)) & \stackrel{\text{def}}{=} & \mathit{Stars}\ (\mathit{inj}\ r\ c\ v :: vs)
\end{array}$$

[1] POSIX lexing acquired its name from the fact that the corresponding rules were described as part of the POSIX specification for Unix-like operating systems [1].

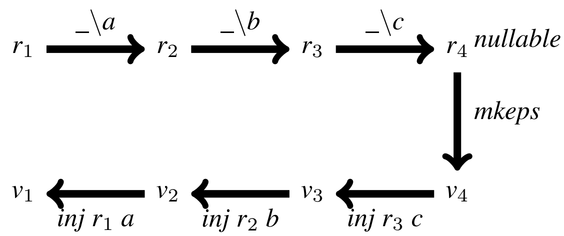

Figure 2 The two phases of the first algorithm by Sulzmann & Lu [12], matching the string $[a, b, c]$. The first phase (the arrows from left to right) is Brzozowski's matcher building successive derivatives. If the last regular expression is *nullable*, then the functions of the second phase are called (the top-down and right-to-left arrows): first *mkeps* calculates a value v_4 witnessing how the empty string has been recognised by r_4. After that the function *inj* "injects back" the characters of the string into the values. The value v_1 is the result of the algorithm representing the POSIX value for this string and regular expression.

The function *mkeps* is run when the last derivative is nullable, that is the string to be matched is in the language of the regular expression. It generates a value for how the last derivative can match the empty string. The injection function then calculates the corresponding value for each intermediate derivative until a value for the original regular expression is generated. Graphically the algorithm by Sulzmann and Lu can be illustrated by the picture in Figure 2 where the path from the left to the right involving *derivatives*/*nullable* is the first phase of the algorithm (calculating successive Brzozowski's derivatives) and *mkeps*/*inj*, the path from right to left, the second phase. The picture above shows the steps required when a regular expression, say r_1, matches the string $[a, b, c]$. The first lexing algorithm by Sulzmann and Lu can be defined as:

$$\begin{array}{lcl} \textit{lexer}\ r\ [] & \stackrel{\text{def}}{=} & \textit{if nullable}\ r\ \textit{then Some}\ (\textit{mkeps}\ r)\ \textit{else None} \\ \textit{lexer}\ r\ (c :: s) & \stackrel{\text{def}}{=} & \textit{case lexer}\ (r\backslash c)\ s\ \textit{of} \\ & & \quad \textit{None} \Rightarrow \textit{None} \\ & & \mid \textit{Some}\ v \Rightarrow \textit{Some}\ (\textit{inj}\ r\ c\ v) \end{array}$$

We have shown in our earlier paper [3] that this algorithm is correct, that is it generates POSIX values. The central property we established relates the derivative operation to the injection function.

▶ **Proposition 3.** *If* $(s,\ r\backslash c) \to v$ *then* $(c :: s,\ r) \to \textit{inj}\ r\ c\ v$.

With this in place we were able to prove:

▶ **Proposition 4.**

(1) $s \notin L\ r$ *if and only if* $\textit{lexer}\ r\ s = \textit{None}$
(2) $s \in L\ r$ *if and only if* $\exists v.\ \textit{lexer}\ r\ s = \textit{Some}\ v \wedge (s, r) \to v$

In fact we have shown that in the success case the generated POSIX value v is unique and in the failure case that there is no POSIX value v that satisfies $(s, r) \to v$. While the algorithm is correct, it is excruciatingly slow in cases where the derivatives grow arbitrarily (recall the example from the Introduction). However it can be used as a convenient reference point for the correctness proof of the second algorithm by Sulzmann and Lu, which we shall describe next.

3 Bitcoded Regular Expressions and Derivatives

In the second part of their paper [12], Sulzmann and Lu describe another algorithm that also generates POSIX values but dispenses with the second phase where characters are injected "back" into values. For this they annotate bitcodes to regular expressions, which we define in Isabelle/HOL as the datatype

$$\begin{array}{lcl}
\textit{breg} & ::= & \textit{ZERO} \quad | \quad \textit{ONE}\ \textit{bs}\\
& | & \textit{CHAR}\ \textit{bs}\ c\\
& | & \textit{ALTs}\ \textit{bs}\ \textit{rs}\\
& | & \textit{SEQ}\ \textit{bs}\ r_1\ r_2\\
& | & \textit{STAR}\ \textit{bs}\ r
\end{array}$$

where bs stands for bitsequences; r, r_1 and r_2 for bitcoded regular expressions; and rs for lists of bitcoded regular expressions. The binary alternative $\textit{ALT}\ bs\ r_1\ r_2$ is just an abbreviation for $\textit{ALTs}\ bs\ [r_1, r_2]$. For bitsequences we use lists made up of the constants Z and S. The idea with bitcoded regular expressions is to incrementally generate the value information (for example *Left* and *Right*) as bitsequences. For this Sulzmann and Lu define a coding function for how values can be coded into bitsequences.

$$\begin{array}{lcl@{\qquad}lcl}
\textit{code}\ (\textit{Empty}) & \stackrel{\text{def}}{=} & [] & \textit{code}\ (\textit{Seq}\ v_1\ v_2) & \stackrel{\text{def}}{=} & \textit{code}\ v_1\ @\ \textit{code}\ v_2\\
\textit{code}\ (\textit{Char}\ c) & \stackrel{\text{def}}{=} & [] & \textit{code}\ (\textit{Stars}\ []) & \stackrel{\text{def}}{=} & [S]\\
\textit{code}\ (\textit{Left}\ v) & \stackrel{\text{def}}{=} & Z :: \textit{code}\ v & \textit{code}\ (\textit{Stars}\ (v :: vs)) & \stackrel{\text{def}}{=} & Z :: \textit{code}\ v\ @\ \textit{code}\ (\textit{Stars}\ vs)\\
\textit{code}\ (\textit{Right}\ v) & \stackrel{\text{def}}{=} & S :: \textit{code}\ v & & &
\end{array}$$

As can be seen, this coding is "lossy" in the sense that we do not record explicitly character values and also not sequence values (for them we just append two bitsequences). However, the different alternatives for *Left*, respectively *Right*, are recorded as Z and S followed by some bitsequence. Similarly, we use Z to indicate if there is still a value coming in the list of *Stars*, whereas S indicates the end of the list. The lossiness makes the process of decoding a bit more involved, but the point is that if we have a regular expression *and* a bitsequence of a corresponding value, then we can always decode the value accurately. The decoding can be defined by using two functions called $\textit{decode}'$ and *decode*:

$$\begin{array}{lcl}
\textit{decode}'\ bs\ (\mathbf{1}) & \stackrel{\text{def}}{=} & (\textit{Empty}, bs)\\
\textit{decode}'\ bs\ (c) & \stackrel{\text{def}}{=} & (\textit{Char}\ c, bs)\\
\textit{decode}'\ (Z :: bs)\ (r_1 + r_2) & \stackrel{\text{def}}{=} & \textit{let}\ (v, bs_1) = \textit{decode}'\ bs\ r_1\ \textit{in}\ (\textit{Left}\ v, bs_1)\\
\textit{decode}'\ (S :: bs)\ (r_1 + r_2) & \stackrel{\text{def}}{=} & \textit{let}\ (v, bs_1) = \textit{decode}'\ bs\ r_2\ \textit{in}\ (\textit{Right}\ v, bs_1)\\
\textit{decode}'\ bs\ (r_1 \cdot r_2) & \stackrel{\text{def}}{=} & \textit{let}\ (v_1, bs_1) = \textit{decode}'\ bs\ r_1\ \textit{in}\\
& & \textit{let}\ (v_2, bs_2) = \textit{decode}'\ bs_1\ r_2\quad \textit{in}\ (\textit{Seq}\ v_1\ v_2, bs_2)\\
\textit{decode}'\ (Z :: bs)\ (r^*) & \stackrel{\text{def}}{=} & (\textit{Stars}\ [], bs)\\
\textit{decode}'\ (S :: bs)\ (r^*) & \stackrel{\text{def}}{=} & \textit{let}\ (v, bs_1) = \textit{decode}'\ bs\ r\ \textit{in}\\
& & \textit{let}\ (\textit{Stars}\ vs, bs_2) = \textit{decode}'\ bs_1\ r^*\quad \textit{in}\ (\textit{Stars}\ v :: vs, bs_2)\\
\\
\textit{decode}\ bs\ r & \stackrel{\text{def}}{=} & \textit{let}\ (v, bs') = \textit{decode}'\ bs\ r\ \textit{in}\\
& & \quad \textit{if}\ bs' = []\ \textit{then}\ \textit{Some}\ v\ \textit{else}\ \textit{None}
\end{array}$$

The function *decode* checks whether all of the bitsequence is consumed and returns the corresponding value as *Some* v; otherwise it fails with *None*. We can establish that for a value v inhabited by a regular expression r, the decoding of its bitsequence never fails.

▶ **Lemma 5.** *If* $\vdash v : r$ *then* $\textit{decode}\ (\textit{code}\ v)\ r = \textit{Some}\ v$.

Proof. This follows from the property that $\textit{decode}'\ ((\textit{code}\ v)\ @\ bs)\ r = (v, bs)$ holds for any bitsequence bs and $\vdash v : r$. This property can be easily proved by induction on $\vdash v : r$. ◀

Sulzmann and Lu define the function *internalise* in order to transform (standard) regular expressions into annotated regular expressions. We write this operation as $r^\uparrow$. This internalisation uses the following *fuse* function.

$$
\begin{array}{lcl}
\textit{fuse}\, bs\, (\textit{ZERO}) & \stackrel{\text{def}}{=} & \textit{ZERO} \\
\textit{fuse}\, bs\, (\textit{ONE}\, bs') & \stackrel{\text{def}}{=} & \textit{ONE}\, (bs\,@\,bs') \\
\textit{fuse}\, bs\, (\textit{CHAR}\, bs'\, c) & \stackrel{\text{def}}{=} & \textit{CHAR}\, (bs\,@\,bs')\, c \\
\textit{fuse}\, bs\, (\textit{ALTs}\, bs'\, rs) & \stackrel{\text{def}}{=} & \textit{ALTs}\, (bs\,@\,bs')\, rs \\
\textit{fuse}\, bs\, (\textit{SEQ}\, bs'\, r_1\, r_2) & \stackrel{\text{def}}{=} & \textit{SEQ}\, (bs\,@\,bs')\, r_1\, r_2 \\
\textit{fuse}\, bs\, (\textit{STAR}\, bs'\, r) & \stackrel{\text{def}}{=} & \textit{STAR}\, (bs\,@\,bs')\, r
\end{array}
$$

A regular expression can then be *internalised* into a bitcoded regular expression as follows:

$$
\begin{array}{lcl}
(\mathbf{0})^\uparrow & \stackrel{\text{def}}{=} & \textit{ZERO} \\
(\mathbf{1})^\uparrow & \stackrel{\text{def}}{=} & \textit{ONE}\, [] \\
(c)^\uparrow & \stackrel{\text{def}}{=} & \textit{CHAR}\, []\, c \\
(r_1 + r_2)^\uparrow & \stackrel{\text{def}}{=} & \textit{ALT}\, []\, (\textit{fuse}\, [Z]\, r_1^\uparrow)\, (\textit{fuse}\, [S]\, r_2^\uparrow) \\
(r_1 \cdot r_2)^\uparrow & \stackrel{\text{def}}{=} & \textit{SEQ}\, []\, r_1^\uparrow\, r_2^\uparrow \\
(r^*)^\uparrow & \stackrel{\text{def}}{=} & \textit{STAR}\, []\, r^\uparrow
\end{array}
$$

There is also an *erase*-function, written $r^\downarrow$, which transforms a bitcoded regular expression into a (standard) regular expression by just erasing the annotated bitsequences. We omit the straightforward definition. For defining the algorithm, we also need the functions *bnullable* and *bmkeps*(s), which are the "lifted" versions of *nullable* and *mkeps* acting on bitcoded regular expressions.

$$
\begin{array}{lcl}
\textit{bnullable}\, (\textit{ZERO}) & \stackrel{\text{def}}{=} & \textit{False} \\
\textit{bnullable}\, (\textit{ONE}\, bs) & \stackrel{\text{def}}{=} & \textit{True} \\
\textit{bnullable}\, (\textit{CHAR}\, bs\, c) & \stackrel{\text{def}}{=} & \textit{False} \\
\textit{bnullable}\, (\textit{ALTs}\, bs\, rs) & \stackrel{\text{def}}{=} & \exists r \in rs.\, \textit{bnullable}\, r \\
\textit{bnullable}\, (\textit{SEQ}\, bs\, r_1\, r_2) & \stackrel{\text{def}}{=} & \textit{bnullable}\, r_1 \wedge \textit{bnullable}\, r_2 \\
\textit{bnullable}\, (\textit{STAR}\, bs\, r) & \stackrel{\text{def}}{=} & \textit{True}
\end{array}
$$

$$
\begin{array}{lcl}
\textit{bmkeps}\, (\textit{ONE}\, bs) & \stackrel{\text{def}}{=} & bs \\
\textit{bmkeps}\, (\textit{ALTs}\, bs\, rs) & \stackrel{\text{def}}{=} & bs\, @\, \textit{bmkepss}\, rs \\
\textit{bmkeps}\, (\textit{SEQ}\, bs\, r_1\, r_2) & \stackrel{\text{def}}{=} & \\
\multicolumn{3}{r}{bs\, @\, \textit{bmkeps}\, r_1\, @\, \textit{bmkeps}\, r_2} \\
\textit{bmkeps}\, (\textit{STAR}\, bs\, r) & \stackrel{\text{def}}{=} & bs\, @\, [S] \\
\textit{bmkepss}\, (r :: rs) & \stackrel{\text{def}}{=} & \textit{if bnullable}\, r \\
 & & \textit{then bmkeps}\, r \\
 & & \textit{else bmkepss}\, rs
\end{array}
$$

The key function in the bitcoded algorithm is the derivative of a bitcoded regular expression. This derivative function calculates the derivative but at the same time also the incremental part of the bitsequences that contribute to constructing a POSIX value.

$$
\begin{array}{lcl}
(\textit{ZERO})\backslash c & \stackrel{\text{def}}{=} & \textit{ZERO} \\
(\textit{ONE}\ bs)\backslash c & \stackrel{\text{def}}{=} & \textit{ZERO} \\
(\textit{CHAR}\ bs\, d)\backslash c & \stackrel{\text{def}}{=} & \textit{if}\ c = d\ \ \textit{then ONE bs else ZERO} \\
(\textit{ALTs}\ bs\ rs)\backslash c & \stackrel{\text{def}}{=} & \textit{ALTs}\, bs\, (\textit{map}\, (\_\backslash c)\ rs) \\
(\textit{SEQ}\ bs\, r_1\, r_2)\backslash c & \stackrel{\text{def}}{=} & \textit{if bnullable}\, r_1 \\
 & & \textit{then ALT}\, bs\, (\textit{SEQ}\, []\, (r_1\backslash c)\, r_2) \\
 & & \qquad (\textit{fuse}\, (\textit{bmkeps}\, r_1)\, (r_2\backslash c)) \\
 & & \textit{else SEQ}\, bs\, (r_1\backslash c)\, r_2 \\
(\textit{STAR}\, bs\, r)\backslash c & \stackrel{\text{def}}{=} & \textit{SEQ}\ bs\, (\textit{fuse}\, [Z] (r\backslash c))\, (\textit{STAR}\, []\, r)
\end{array}
$$

This function can also be extended to strings, written $r\backslash s$, just like the standard derivative. We omit the details. Finally we can define Sulzmann and Lu's bitcoded lexer, which we call *blexer*:

$$
\begin{array}{lcl}
\textit{blexer}\ r\ s & \stackrel{\text{def}}{=} & \textit{let}\ r_{der} = (r^\uparrow)\backslash s\ \textit{in} \\
 & & \quad \textit{if bnullable}(r_{der})\ \ \textit{then decode}\, (\textit{bmkeps}\, r_{der})\, r\ \ \textit{else None}
\end{array}
$$

This bitcoded lexer first internalises the regular expression r and then builds the bitcoded derivative according to s. If the derivative is (b)nullable the string is in the language of r and it extracts the bitsequence using the *bmkeps* function. Finally it decodes the bitsequence into a value. If the derivative is *not* nullable, then *None* is returned. We can show that this way of calculating a value generates the same result as *lexer*.

Before we can proceed we need to define a helper function, called *retrieve*, which Sulzmann and Lu introduced for the correctness proof.

$$
\begin{array}{lcl}
\textit{retrieve}\ (\textit{ONE}\ \textit{bs})\ (\textit{Empty}) & \stackrel{\text{def}}{=} & \textit{bs}\\
\textit{retrieve}\ (\textit{CHAR}\ \textit{bs}\ c)\ (\textit{Char}\ d) & \stackrel{\text{def}}{=} & \textit{bs}\\
\textit{retrieve}\ (\textit{ALTs}\ \textit{bs}\ [r])\ v & \stackrel{\text{def}}{=} & \textit{bs}\ @\ \textit{retrieve}\ r\ v\\
\textit{retrieve}\ (\textit{ALTs}\ \textit{bs}\ (r :: \textit{rs}))\ (\textit{Left}\ v) & \stackrel{\text{def}}{=} & \textit{bs}\ @\ \textit{retrieve}\ r\ v\\
\textit{retrieve}\ (\textit{ALTs}\ \textit{bs}\ (r :: \textit{rs}))\ (\textit{Right}\ v) & \stackrel{\text{def}}{=} & \textit{bs}\ @\ \textit{retrieve}\ (\textit{ALTs}\ []\ \textit{rs})\ v\\
\textit{retrieve}\ (\textit{SEQ}\ \textit{bs}\ r_1\ r_2)\ (\textit{Seq}\ v_1\ v_2) & \stackrel{\text{def}}{=} & \textit{bs}\ @\ \textit{retrieve}\ r_1\ v_1\ @\ \textit{retrieve}\ r_2\ v_2\\
\textit{retrieve}\ (\textit{STAR}\ \textit{bs}\ r)\ (\textit{Stars}\ []) & \stackrel{\text{def}}{=} & \textit{bs}\ @\ [S]\\
\textit{retrieve}\ (\textit{STAR}\ \textit{bs}\ r)\ (\textit{Stars}\ (v :: \textit{vs})) & \stackrel{\text{def}}{=} & \textit{bs}\ @\ [Z]\ @\ \textit{retrieve}\ r\ v\ @\ \textit{retrieve}\ (\textit{STAR}\ []\ r)\ (\textit{Stars}\ \textit{vs})
\end{array}
$$

The idea behind this function is to retrieve a possibly partial bitsequence from a bitcoded regular expression, where the retrieval is guided by a value. For example if the value is $Left$ then we descend into the left-hand side of an alternative in order to assemble the bitcode. Similarly for $Right$. The property we can show is that for a given v and r with $\vdash v : r$, the retrieved bitsequence from the internalised regular expression is equal to the bitcoded version of v.

▶ **Lemma 6.** *If* $\vdash v : r$ *then* $\textit{code}\ v = \textit{retrieve}\ (r^{\uparrow})\ v$.

We also need some auxiliary facts about how the bitcoded operations relate to the "standard" operations on regular expressions. For example if we build a bitcoded derivative and erase the result, this is the same as if we first erase the bitcoded regular expression and then perform the "standard" derivative operation.

▶ **Lemma 7.**

(1) $(r\backslash s)^{\downarrow} = (r^{\downarrow})\backslash s$

(2) $\textit{bnullable}(r)$ *iff* $\textit{nullable}(r^{\downarrow})$

(3) $\textit{bmkeps}(r) = \textit{retrieve}\ r\ (\textit{mkeps}\ (r^{\downarrow}))$ *provided* $\textit{nullable}(r^{\downarrow})$.

Proof. All properties are by induction on annotated regular expressions. There are no interesting cases. ◀

The only difficulty left for the correctness proof is that the bitcoded algorithm has only a "forward phase" where POSIX values are generated incrementally. We can achieve the same effect with *lexer* (which has two phases) by stacking up injection functions during the forward phase. An auxiliary function, called *flex*, allows us to recast the rules of $lexer$ in terms of a single phase and stacked up injection functions.

$$
\begin{array}{lcl}
\textit{flex}\ r\ f\ [] & \stackrel{\text{def}}{=} & f\\
\textit{flex}\ r\ f\ (c{::}s) & \stackrel{\text{def}}{=} & \textit{flex}\ (r\backslash c)\ (\lambda v.\ f\ (\textit{inj}\ r\ c\ v))\ s
\end{array}
$$

The point of this function is that when reaching the end of the string, we just need to apply the stacked up injection functions to the value generated by *mkeps*. Using this function we can recast the success case in *lexer* as follows:

▶ **Proposition 8.** *If* $\textit{lexer}\; r\; s = \textit{Some}\; v$ *then* $v = \textit{flex}\; r\; \textit{id}\; s\; (\textit{mkeps}(r\backslash s))$.

Note we did not redefine *lexer*, we just established that the value generated by *lexer* can also be obtained by a different method. While this different method is not efficient (we essentially need to traverse the string s twice, once for building the derivative $r\backslash s$ and another time for stacking up injection functions using *flex*), it helps us with proving that incrementally building up values in *blexer* generates the same result.

This brings us to our main lemma in this section: if we calculate a derivative, say $r\backslash s$, and have a value, say v, inhabited by this derivative, then we can produce the result *lexer* generates by applying this value to the stacked-up injection functions that *flex* assembles. The lemma establishes that this is the same value as if we build the annotated derivative $r^{\uparrow}\backslash s$ and then retrieve the corresponding bitcoded version, followed by a decoding step.

▶ **Lemma 9** (Main Lemma). *If* $\vdash v : r\backslash s$ *then*

$$\textit{Some}\; (\textit{flex}\; r\; \textit{id}\; s\; v) = \textit{decode}(\textit{retrieve}\; (r^{\uparrow}\backslash s)\; v)\; r$$

Proof. This can be proved by induction on s and generalising over v. The interesting point is that we need to prove this in the reverse direction for s. This means instead of cases $[]$ and $c :: s$, we have cases $[]$ and $s\,@\,[c]$ where we unravel the string from the back.[2]

The case for $[]$ is routine using Lemmas 5 and 6. In the case $s\,@\,[c]$, we can infer from the assumption that $\vdash v : (r\backslash s)\backslash c$ holds. Hence by Prop. 3 we know that (*) $\vdash \textit{inj}\; (r\backslash s)\; c\; v : r\backslash s$ holds too. By definition of *flex* we can unfold the left-hand side to be

$$\textit{Some}\; (\textit{flex}\; r\; \textit{id}\; (s\,@\,[c])\; v) = \textit{Some}\; (\textit{flex}\; r\; \textit{id}\; s\; (\textit{inj}\; (r\backslash s)\; c\; v))$$

By induction hypothesis and (*) we can rewrite the right-hand side to

$$\textit{decode}\; (\textit{retrieve}\; (r^{\uparrow}\backslash s)\; (\textit{inj}\; (r\backslash s)\; c\; v))\; r$$

which is equal to $\textit{decode}\; (\textit{retrieve}\; (r^{\uparrow}\backslash (s\,@\,[c]))\; v)\; r$ as required. The last rewrite step is possible because we generalised over v in our induction. ◀

With this lemma in place, we can prove the correctness of *blexer*—it indeed produces the same result as *lexer*.

▶ **Theorem 10.** $\textit{lexer}\; r\; s = \textit{blexer}\; r\; s$

Proof. We can first expand both sides using Prop. 8 and the definition of *blexer*. This gives us two *if*-statements, which we need to show to be equal. By Lemma 7*(2)* we know the *if*-tests coincide:

$$\textit{bnullable}(r^{\uparrow}\backslash s)\;\; \textit{iff}\;\; \textit{nullable}(r\backslash s)$$

For the *if*-branch suppose $r_d \stackrel{\text{def}}{=} r^{\uparrow}\backslash s$ and $d \stackrel{\text{def}}{=} r\backslash s$. We have (*) *nullable* d. We can then show by Lemma 7*(3)* that

$$\textit{decode}(\textit{bmkeps}\; r_d)\; r = \textit{decode}(\textit{retrieve}\; r_d\; (\textit{mkeps}\; d))\; r$$

where the right-hand side is equal to $\textit{Some}\; (\textit{flex}\; r\; \textit{id}\; s\; (\textit{mkeps}\; d))$ by Lemma 9 (we know $\vdash \textit{mkeps}\; d : d$ by (*)). This shows the *if*-branches return the same value. In the *else*-branches both *lexer* and *blexer* return *None*. Therefore we can conclude the proof. ◀

This establishes that the bitcoded algorithm by Sulzmann and Lu *without* simplification produces correct results. This was only conjectured by Sulzmann and Lu in their paper [12]. The next step is to add simplifications.

---

[2] Isabelle/HOL provides an induction principle for this way of performing the induction.

4 Simplification

Derivatives as calculated by Brzozowski's method are usually more complex regular expressions than the initial one; the result is that derivative-based matching and lexing algorithms are often abysmally slow if the "growth problem" is not addressed. As Sulzmann and Lu wrote, various optimisations are possible, such as the simplifications $\mathbf{0}\,r \Rightarrow \mathbf{0}$, $\mathbf{1}\,r \Rightarrow r$, $\mathbf{0} + r \Rightarrow r$ and $r + r \Rightarrow r$. While these simplifications can considerably speed up the two algorithms in many cases, they do not solve fundamentally the growth problem with derivatives. To see this let us return to the example from the Introduction that shows the derivatives for $(a + aa)^*$. If we delete in the 3rd step all $\mathbf{0}s$ and $\mathbf{1}$s according to the simplification rules shown above we obtain

$$(a + aa)^* \quad \xrightarrow{\_\backslash[a,a,a]} \quad \underbrace{(\mathbf{1} + \mathbf{1}a) \cdot (a + aa)^*}_{r} + ((a + aa)^* + \underbrace{(\mathbf{1} + \mathbf{1}a) \cdot (a + aa)^*}_{r}) \tag{1}$$

This is a simpler derivative, but unfortunately we cannot make any further simplifications. This is a problem because the outermost alternatives contains two copies of the same regular expression (underlined with r). These copies will spawn new copies in later derivative steps and they in turn even more copies. This destroys any hope of taming the size of the derivatives. But the second copy of r in (1) will never contribute to a value, because POSIX lexing will always prefer matching a string with the first copy. So it could be safely removed without affecting the correctness of the algorithm. The dilemma with the simple-minded simplification rules above is that the rule $r + r \Rightarrow r$ will never be applicable because as can be seen in this example the regular expressions are not next to each other but separated by another regular expression.

But here is where Sulzmann and Lu's representation of generalised alternatives in the bitcoded algorithm shines: in *ALTs bs rs* we can define a more aggressive simplification by recursively simplifying all regular expressions in *rs* and then analyse the resulting list and remove any duplicates. Another advantage with the bitsequences in bitcoded regular expressions is that they can be easily modified such that simplification does not interfere with the value constructions. For example we can "flatten", or de-nest, *ALTs* as follows

$$\textit{ALTs}\ \textit{bs}_1\ (\textit{ALTs}\ \textit{bs}_2\ \textit{rs}_2 :: \textit{rs}_1) \quad \xrightarrow{bsimp} \quad \textit{ALTs}\ \textit{bs}_1\ (\textit{map}\ (\textit{fuse}\ \textit{bs}_2)\ \textit{rs}_2 :: \textit{rs}_1)$$

where we just need to fuse the bitsequence that has accumulated in bs_2 to the alternatives in rs_2. As we shall show below this will ensure that the correct value corresponding to the original (unsimplified) regular expression can still be extracted.

However there is one problem with the definition for the more aggressive simplification rules described by Sulzmann and Lu. Recasting their definition with our syntax they define the step of removing duplicates as

$$\textit{bsimp}\ (\textit{ALTs}\ \textit{bs}\ \textit{rs}) \stackrel{\text{def}}{=} \textit{ALTs}\ \textit{bs}\ (\textit{nub}\ (\textit{map}\ \textit{bsimp}\ \textit{rs}))$$

where they first recursively simplify the regular expressions in *rs* (using *map*) and then use Haskell's *nub*-function to remove potential duplicates. While this makes sense when considering the example shown in (1), *nub* is the inappropriate function in the case of bitcoded regular expressions. The reason is that in general the elements in *rs* will have a different annotated bitsequence and in this way *nub* will never find a duplicate to be removed. One correct way to handle this situation is to first *erase* the regular expressions when comparing potential duplicates. This is inspired by Scala's list functions of the form *distinctBy rs f acc* where a function is applied first before two elements are compared. We define this function in Isabelle/HOL as

$$\begin{aligned}
&\textit{distinctBy}\ []\ f\ \textit{acc} &&\stackrel{\text{def}}{=} []\\
&\textit{distinctBy}\ (x :: xs)\ f\ \textit{acc} &&\stackrel{\text{def}}{=} \textit{if}\ f\,x \in \textit{acc}\ \textit{then}\ \textit{distinctBy}\ xs\ f\ \textit{acc}\ \textit{else}\ x :: \textit{distinctBy}\ xs\ f\ (\{f\,x\} \cup \textit{acc})
\end{aligned}$$

where we scan the list from left to right (because we have to remove later copies). In *distinctBy*, f is a function and *acc* is an accumulator for regular expressions—essentially a set of regular expressions that we have already seen while scanning the list. Therefore we delete an element, say x, from the list provided $f\,x$ is already in the accumulator; otherwise we keep x and scan the rest of the list but add $f\,x$ as another "seen" element to *acc*. We will use *distinctBy* where f is the erase function, $\_^{\downarrow}$, that deletes bitsequences from bitcoded regular expressions. This is clearly a computationally more expensive operation, than *nub*, but is needed in order to make the removal of unnecessary copies to work properly.

Our simplification function depends on three helper functions, one is called *flts* and analyses lists of regular expressions coming from alternatives. It is defined as follows:

$$\begin{array}{lcl} \mathit{flts}\ [] & \stackrel{\text{def}}{=} & [] \\ \mathit{flts}\ (\mathit{ZERO} :: \mathit{rs}) & \stackrel{\text{def}}{=} & \mathit{flts}\ \mathit{rs} \\ \mathit{flts}\ (\mathit{ALTs}\ \mathit{bs}'\ \mathit{rs}' :: \mathit{rs}) & \stackrel{\text{def}}{=} & \mathit{map}\ (\mathit{fuse}\ \mathit{bs}')\ \mathit{rs}'\ @\ \mathit{flts}\ \mathit{rs} \end{array}$$

The second clause of *flts* removes all instances of *ZERO* in alternatives and the third "spills" out nested alternatives (but retaining the bitsequence bs' accumulated in the inner alternative). There are some corner cases to be considered when the resulting list inside an alternative is empty or a singleton list. We take care of those cases in the *bsimpALTs* function; similarly we define a helper function that simplifies sequences according to the usual rules about *ZERO*s and *ONE*s:

$$\begin{array}{lcl} \mathit{bsimpALTs}\ \mathit{bs}\ [] & \stackrel{\text{def}}{=} & \mathit{ZERO} \\ \mathit{bsimpALTs}\ \mathit{bs}\ [r] & \stackrel{\text{def}}{=} & \mathit{fuse}\ \mathit{bs}\ r \\ \mathit{bsimpALTs}\ \mathit{bs}\ \mathit{rs} & \stackrel{\text{def}}{=} & \mathit{ALTs}\ \mathit{bs}\ \mathit{rs} \end{array} \qquad \begin{array}{lcl} \mathit{bsimpSEQ}\ \mathit{bs}\ \_\ \mathit{ZERO} & \stackrel{\text{def}}{=} & \mathit{ZERO} \\ \mathit{bsimpSEQ}\ \mathit{bs}\ \mathit{ZERO}\ \_ & \stackrel{\text{def}}{=} & \mathit{ZERO} \\ \mathit{bsimpSEQ}\ \mathit{bs}_1\ (\mathit{ONE}\ \mathit{bs}_2)\ r_2 & \stackrel{\text{def}}{=} & \mathit{fuse}\ (\mathit{bs}_1\ @\ \mathit{bs}_2)\ r_2 \\ \mathit{bsimpSEQ}\ \mathit{bs}\ r_1\ r_2 & \stackrel{\text{def}}{=} & \mathit{SEQ}\ \mathit{bs}\ r_1\ r_2 \end{array}$$

With this in place we can define our simplification function as

$$\begin{array}{lcl} \mathit{bsimp}\ (\mathit{SEQ}\ \mathit{bs}\ r_1\ r_2) & \stackrel{\text{def}}{=} & \mathit{bsimpSEQ}\ \mathit{bs}\ (\mathit{bsimp}\ r_1)\ (\mathit{bsimp}\ r_2) \\ \mathit{bsimp}\ (\mathit{ALTs}\ \mathit{bs}\ \mathit{rs}) & \stackrel{\text{def}}{=} & \mathit{bsimpALT}\ \mathit{bs}\ (\mathit{distinctBy}\ (\mathit{flts}\ (\mathit{map}\ \mathit{bsimp}\ \mathit{rs}))\ \mathit{erase}\ \varnothing) \\ \mathit{bsimp}\ r & \stackrel{\text{def}}{=} & r \end{array}$$

As far as we can see, our recursive function *bsimp* simplifies regular expressions as intended by Sulzmann and Lu. There is no point in applying the *bsimp* function repeatedly (like the simplification in their paper which needs to be applied until a fixpoint is reached) because we can show that *bsimp* is idempotent, that is

▶ **Proposition 11.** $\mathit{bsimp}\ (\mathit{bsimp}\ r) = \mathit{bsimp}\ r$

This can be proved by induction on r but requires a detailed analysis that the de-nesting of alternatives always results in a flat list of regular expressions. We omit the details since it does not concern the correctness proof.

Next we can include simplification after each derivative step leading to the following notion of bitcoded derivatives:

$$r\backslash_{bsimp}\ [] \stackrel{\text{def}}{=} r \qquad r\backslash_{bsimp}\ (c :: s) \stackrel{\text{def}}{=} \mathit{bsimp}\ (r\backslash c)\backslash_{bsimp}\ s$$

and use it in the improved lexing algorithm defined as

$$\begin{array}{lcl} \mathit{blexer}^{+}\ r\ s & \stackrel{\text{def}}{=} & \mathit{let}\ r_{der} = (r^{\uparrow})\backslash_{bsimp}\ s\ \mathit{in} \\ & & \quad \mathit{if}\ \mathit{bnullable}(r_{der})\ \ \mathit{then}\ \mathit{decode}\ (\mathit{bmkeps}\ r_{der})\ r\ \ \mathit{else}\ \mathit{None} \end{array}$$

The remaining task is to show that *blexer* and $blexer^+$ generate the same answers.

When we first attempted this proof we encountered a problem with the idea in Sulzmann and Lu's paper where the argument seems to be to appeal again to the *retrieve*-function defined for the unsimplified version of the algorithm. But this does not work, because desirable properties such as

$$retrieve\ r\ v = retrieve\ (bsimp\ r)\ v$$

do not hold under simplification—this property essentially purports that we can retrieve the same value from a simplified version of the regular expression. To start with *retrieve* depends on the fact that the value v correspond to the structure of the regular expressions—but the whole point of simplification is to "destroy" this structure by making the regular expression simpler. To see this consider the regular expression $r = r' + 0$ and a corresponding value $v = Left\ v'$. If we annotate bitcodes to r, then we can use *retrieve* and v in order to extract a corresponding bitsequence. The reason that this works is that r is an alternative regular expression and v a corresponding value. However, if we simplify r, then v does not correspond to the shape of the regular expression anymore. So unless one can somehow synchronise the change in the simplified regular expressions with the original POSIX value, there is no hope of appealing to *retrieve* in the correctness argument for $blexer^+$.

We found it more helpful to introduce the rewriting systems shown in Figure 3. The idea is to generate simplified regular expressions in small steps (unlike the *bsimp*-function which does the same in a big step), and show that each of the small steps preserves the bitcodes that lead to the final POSIX value. The rewrite system is organised such that $\leadsto$ is for bitcoded regular expressions and $\overset{s}{\leadsto}$ for lists of bitcoded regular expressions. The former essentially implements the simplifications of *bsimpSEQ* and *flts*; while the latter implements the simplifications in *bsimpALTs*. We can show that any bitcoded regular expression reduces in zero or more steps to the simplified regular expression generated by *bsimp*:

▶ **Lemma 12.** $r \leadsto^* bsimp\ r$

Proof. By induction on r. For this we can use the properties $rs \overset{s}{\leadsto}^* flts\ rs$ and $rs \overset{s}{\leadsto}^* distinctBy\ rs\ erase\ \varnothing$. The latter uses repeated applications of the LD rule which allows the removal of duplicates that can recognise the same strings. ◀

We can show that this rewrite system preserves *bnullable*, that is simplification, essentially, does not affect nullability:

▶ **Lemma 13.** *If* $r_1 \leadsto r_2$ *then* $bnullable\ r_1 = bnullable\ r_2$.

Proof. Straightforward mutual induction on the definition of $\leadsto$ and $\overset{s}{\leadsto}$. The only interesting case is the rule LD where the property holds since by the side-conditions of that rule the empty string will be in both $L\ (rs_a\ @\ [r_1]\ @\ rs_b\ @\ [r_2]\ @\ rs_c)$ and $L\ (rs_a\ @\ [r_1]\ @\ rs_b\ @\ rs_c)$. ◀

From this, we can show that *bmkeps* will produce the same bitsequence as long as one of the bitcoded regular expressions in $\leadsto$ is nullable (this lemma establishes the missing fact we were not able to establish using *retrieve*, as suggested in the paper by Sulzmann and Lu).

▶ **Lemma 14.** *If* $r_1 \leadsto r_2$ *and* $bnullable\ r_1$ *then* $bmkeps\ r_1 = bmkeps\ r_2$.

Proof. By straightforward mutual induction on the definition of $\leadsto$ and $\overset{s}{\leadsto}$. Again the only interesting case is the rule LD where we need to ensure that

$$bmkeps\ (rs_a\ @\ [r_1]\ @\ rs_b\ @\ [r_2]\ @\ rs_c) = bmkeps\ (rs_a\ @\ [r_1]\ @\ rs_b\ @\ rs_c)$$

holds. This is indeed the case because according to the POSIX rules the generated bitsequence is determined by the first alternative that can match the string (in this case being nullable). ◀

$$\frac{}{(SEQ\ bs\ ZERO\ r_2) \rightsquigarrow (ZERO)}S\mathbf{0}_l \qquad \frac{}{(SEQ\ bs\ r_1\ ZERO) \rightsquigarrow (ZERO)}S\mathbf{0}_r$$

$$\frac{}{(SEQ\ bs_1\ (ONE\ bs_2)\ r) \rightsquigarrow fuse\ (bs_1 @ bs_2)\ r}S\mathbf{1}$$

$$\frac{r_1 \rightsquigarrow r_2}{(SEQ\ bs\ r_1\ r_3) \rightsquigarrow (SEQ\ bs\ r_2\ r_3)}\text{SL} \qquad \frac{r_3 \rightsquigarrow r_4}{(SEQ\ bs\ r_1\ r_3) \rightsquigarrow (SEQ\ bs\ r_1\ r_4)}\text{SR}$$

$$\frac{}{(ALTs\ bs\ []) \rightsquigarrow (ZERO)}A0 \qquad \frac{}{(ALTs\ bs\ [r]) \rightsquigarrow fuse\ bs\ r}A1$$

$$\frac{rs_1 \overset{s}{\rightsquigarrow} rs_2}{(ALTs\ bs\ rs_1) \rightsquigarrow (ALTs\ bs\ rs_2)}AL$$

$$\frac{rs_1 \overset{s}{\rightsquigarrow} rs_2}{r :: rs_1 \overset{s}{\rightsquigarrow} r :: rs_2}LH \qquad \frac{r_1 \rightsquigarrow r_2}{r_1 :: rs \overset{s}{\rightsquigarrow} r_2 :: rs}LT$$

$$\frac{}{ZERO :: rs \overset{s}{\rightsquigarrow} rs}L\mathbf{0} \qquad \frac{}{ALTs\ bs\ rs_1 :: rs_2 \overset{s}{\rightsquigarrow} (map\ (fuse\ bs)\ rs_1 @ rs_2)}LS$$

$$\frac{L\ (r_2{}^{\downarrow}) \subseteq L\ (r_1{}^{\downarrow})}{(rs_1 @ [r_1] @ rs_2 @ [r_2] @ rs_3) \overset{s}{\rightsquigarrow} (rs_1 @ [r_1] @ rs_2 @ rs_3)}LD$$

Figure 3 The rewrite rules that generate simplified regular expressions in small steps: $r_1 \rightsquigarrow r_2$ is for bitcoded regular expressions and $rs_1 \rightsquigarrow^* rs_2$ for *lists* of bitcoded regular expressions. Interesting is the LD rule that allows copies of regular expressions be removed provided a regular expression earlier in the list can match the same strings.

Crucial is also the fact that derivative steps and simplification steps can be interleaved, which is shown by the fact that $\rightsquigarrow$ is preserved under derivatives.

▶ **Lemma 15.** *If* $r_1 \rightsquigarrow r_2$ *then* $r_1\backslash c \rightsquigarrow^* r_2\backslash c$.

Proof. By straightforward mutual induction on the definition of $\rightsquigarrow$ and $\overset{s}{\rightsquigarrow}$. The case for LD holds because $L\ ((r_2\backslash c)^{\downarrow}) \subseteq L\ ((r_1\backslash c)^{\downarrow})$ if and only if $L\ (r_2{}^{\downarrow}) \subseteq L\ (r_1{}^{\downarrow})$. ◀

Using this fact together with Lemma 12 allows us to prove the central lemma that the unsimplified derivative (with a string s) reduces to the simplified derivative (with the same string).

▶ **Lemma 16.** $r\backslash s \rightsquigarrow^* r\backslash_{bsimp}\ s$

Proof. By reverse induction on s generalising over r. ◀

With these lemmas in place we can finally establish that *blexer*$^+$ and *blexer* generate the same value, and using Theorem 10 from the previous section that this value is indeed the POSIX value.

▶ **Theorem 17.** $blexer\ r\ s = blexer^+\ r\ s$

Proof. By unfolding the definitions and using Lemmas 16 and 14. ◀

This completes the correctness proof for the second POSIX lexing algorithm by Sulzmann and Lu. The interesting point of this algorithm is that the sizes of derivatives do not grow arbitrarily, which we shall show next.

5 Finiteness of Derivatives

In this section let us sketch our argument for why the size of the simplified derivatives with the aggressive simplification function is finite. Suppose we have a size function for bitcoded regular expressions, written $|r|$, which counts the number of nodes if we regard r as a tree (we omit the precise definition). For this we show that for every r there exists a bound N such that

$$\forall s.\ |r \backslash_{bsimp}\, s| < N$$

We prove this by induction on r. The base cases for *ZERO*, *ONE bs* and *CHAR bs c* are straightforward. The interesting case is for sequences of the form *SEQ bs* r_1 r_2. In this case our induction hypotheses state $\forall s.\ |r_1 \backslash_{bsimp}\, s| < N_1$ and $\forall s.\ |r_2 \backslash_{bsimp}\, s| < N_2$. We can reason as follows

$$\begin{array}{rll}
& |(\mathit{SEQ}\ bs\ r_1\ r_2) \backslash_{bsimp}\, s| & \\
= & |\mathit{bsimp}(\mathit{ALTs}\ bs\ ((r_1 \backslash_{bsimp}\, s) \cdot r_2) :: [r_2 \backslash_{bsimp}\, s' \mid s' \in \mathit{Suffix}(s)])| & (1)\\
\leq & |\mathit{distinctBy}\,(\mathit{flts}\,((r_1 \backslash_{bsimp}\, s) \cdot r_2) :: [r_2 \backslash_{bsimp}\, s' \mid s' \in \mathit{Suffix}(s)])| + 1 & (2)\\
\leq & |(r_1 \backslash_{bsimp}\, s) \cdot r_2| + |\mathit{distinctBy}\,(\mathit{flts}\,[r_2 \backslash_{bsimp}\, s' \mid s' \in \mathit{Suffix}(s)])| + 1 & (3)\\
\leq & N_1 + |r_2| + 2 + |\mathit{distinctBy}\,(\mathit{flts}\,[r_2 \backslash_{bsimp}\, s' \mid s' \in \mathit{Suffix}(s)])| & (4)\\
\leq & N_1 + |r_2| + 2 + l_{N_2} * N_2 & (5)
\end{array}$$

where in (1) the $\mathit{Suffix}(s')$ are the suffixes where $r_1 \backslash_{bsimp}\, s''$ is nullable for $s = s'' @ s'$. In (3) we know that $|(r_1 \backslash_{bsimp}\, s) \cdot r_2|$ is bounded by $N_1 + |r_2|$. In (5) we know the list comprehension contains only regular expressions of size smaller than N_2. The list length after *distinctBy* is bounded by a number, which we call l_{N_2}. It stands for the number of distinct regular expressions with a maximum size N_2 (there can only be finitely many of them). We reason similarly in the *Star*-case.

Clearly we give in this finiteness argument (Step (5)) a very loose bound that is far from the actual bound we can expect. We can do better than this, but this does not improve the finiteness property we are proving. If we are interested in a polynomial bound, one would hope to obtain a similar tight bound as for partial derivatives introduced by Antimirov [2]. After all the idea with *distinctBy* is to maintain a "set" of alternatives (like the sets in partial derivatives). Unfortunately to obtain the exact same bound would mean we need to introduce simplifications such as

$$(r_1 + r_2) \cdot r_3 \longrightarrow (r_1 \cdot r_3) + (r_2 \cdot r_3)$$

which exist for partial derivatives. However, if we introduce them in our setting we would lose the POSIX property of our calculated values. We leave better bounds for future work.

6 Conclusion

We set out in this work to prove in Isabelle/HOL the correctness of the second POSIX lexing algorithm by Sulzmann and Lu [12]. This follows earlier work where we established the correctness of the first algorithm [3]. In the earlier work we needed to introduce our own specification about what POSIX values are, because the informal definition given by Sulzmann and Lu did not stand up to a formal proof. Also for the second algorithm we needed to introduce our own definitions and proof ideas in order to establish the correctness. Our interest in the second algorithm lies in the fact that by using bitcoded regular expressions and an aggressive simplification method there is a chance that the the derivatives can be kept universally small (we established in this paper that they can be kept finite for any string). This is important if one is after an efficient POSIX lexing algorithm.

Having proved the correctness of the POSIX lexing algorithm, which lessons have we learned? Well, we feel this is a very good example where formal proofs give further insight into the matter at hand. For example it is very hard to see a problem with *nub* vs *distinctBy* with only experimental data—one would still see the correct result but find that simplification does not simplify in well-chosen, but not obscure, examples. We found that from an implementation point-of-view it is really important to have the formal proofs of the corresponding properties at hand. We have also developed a healthy suspicion when experimental data is used to back up efficiency claims. For example Sulzmann and Lu write about their equivalent of $blexer^+$ "...we can incrementally compute bitcoded parse trees in linear time in the size of the input" [12, Page 14]. Given the growth of the derivatives in some cases even after aggressive simplification, this is a hard to believe fact. A similar claim about a theoretical runtime of $O(n^2)$ is made for the Verbatim lexer, which calculates tokens according to POSIX rules [6]. For this it uses Brzozowski's derivatives . They write: "The results of our empirical tests [..] confirm that Verbatim has $O(n^2)$ time complexity." [6, Section VII]. While their correctness proof for Verbatim is formalised in Coq, the claim about the runtime complexity is only supported by some emperical evidence. In the context of our observation with the "growth problem" of derivatives, we tried out their extracted OCaml code with the example $(a + aa)^*$ as a single lexing rule, and it took for us around 5 minutes to tokenise a string of 40 a's and that increased to approximately 19 minutes when the string is 50 a's long. Given that derivatives are not simplified in the Verbatim lexer, such numbers are not surprising. Clearly our result of having finite derivatives might sound rather weak in this context but we think such effeciency claims really require further scrutiny.

Our Isabelle/HOL code is available under https://github.com/urbanchr/posix.

References

1 The Open Group Base Specification Issue 6 IEEE Std 1003.1 2004 Edition, 2004. http://pubs.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html.

2 V. Antimirov. Partial Derivatives of Regular Expressions and Finite Automata Constructions. *Theoretical Computer Science*, 155:291–319, 1995.

3 F. Ausaf, R. Dyckhoff, and C. Urban. POSIX Lexing with Derivatives of Regular Expressions (Proof Pearl). In *Proc. of the 7th International Conference on Interactive Theorem Proving (ITP)*, volume 9807 of *LNCS*, pages 69–86, 2016.

4 J. A. Brzozowski. Derivatives of Regular Expressions. *Journal of the ACM*, 11(4):481–494, 1964.

5 T. Coquand and V. Siles. A Decision Procedure for Regular Expression Equivalence in Type Theory. In *Proc. of the 1st International Conference on Certified Programs and Proofs (CPP)*, volume 7086 of *LNCS*, pages 119–134, 2011.

6 D. Egolf, S. Lasser, and K. Fisher. Verbatim: A Verified Lexer Generator. In *2021 IEEE Security and Privacy Workshops (SPW)*, pages 92–100, 2021.

7 A. Krauss and T. Nipkow. Proof Pearl: Regular Expression Equivalence and Relation Algebra. *Journal of Automated Reasoning*, 49:95–106, 2012.

8 C. Kuklewicz. Regex Posix. https://wiki.haskell.org/Regex_Posix.

9 S. Okui and T. Suzuki. Disambiguation in Regular Expression Matching via Position Automata with Augmented Transitions. In *Proc. of the 15th International Conference on Implementation and Application of Automata (CIAA)*, volume 6482 of *LNCS*, pages 231–240, 2010.

10 S. Owens and K. Slind. Adapting Functional Programs to Higher Order Logic. *Higher-Order and Symbolic Computation*, 21(4):377–409, 2008.

11 R. Ribeiro and A. Du Bois. Certified Bit-Coded Regular Expression Parsing. In *Proc. of the 21st Brazilian Symposium on Programming Languages*, New York, NY, USA, 2017. Association for Computing Machinery.

12 M. Sulzmann and K. Lu. POSIX Regular Expression Parsing with Derivatives. In *Proc. of the 12th International Conference on Functional and Logic Programming (FLOPS)*, volume 8475 of *LNCS*, pages 203–220, 2014.

13 S. Vansummeren. Type Inference for Unique Pattern Matching. *ACM Transactions on Programming Languages and Systems*, 28(3):389–428, 2006.