Progress Report - 9 Months

Meshal Binnasban

September 23, 2025

Abstract

This is the nine-month progress report for my PhD at King's College London. It summarises the motivation, challenges, and progress made in developing a regular expression matcher based on marks with the goal of POSIX value extraction. The report first reviews the derivative-based approach to regular expression matching and a possible correspondence between both approaches —a connection we aim to investigate in the future. It also reviews the marked algorithms of Fischer et al. and Asperti et al., which provide matchers but do not support value extraction, which we are after. Our work developed several versions of the marked algorithm in Scala aimed at extracting POSIX values. During this process, challenges in handling certain cases led us to refine the algorithm through successive versions. Future directions include extending the matcher to additional operators, refining disambiguation for repetitions, and formally proving correctness.

Synopsis

This research investigates the marked approach–based method for regular expression matching, with the goal of extending it to provide POSIX value extraction.

Derivative-based methods, though elegant, suffer from severe size explosion and remain sensitive to syntactic form. Sulzmann and Lu [9] extended derivatives with bitcodes to record lexing information, but the size problem persists. The marked approach, described in the works of Fischer et al. [4] and Asperti et al. [2], offers an alternative by propagating marks directly through the expression. These works provide matchers only, without value extraction.

Our work explores how the marked approach can be extended to recover POSIX values. We have implemented several versions of the marked algorithm in Scala, making use of bitcoded annotations to track lexing values. Challenges in handling constructs such as sequences and repetitions required refining the algorithm through successive versions, each improving on the last. Large-scale testing against a derivative-based reference matcher has been used to confirm correctness and uncover edge cases.

The long-term aim of the project is to establish a marked approach–based matcher that consistently yields the POSIX value for all regular expressions, and to formally prove its correctness. Future work also includes extending the matcher to additional operators such as intersection and negation.

Contents

1 Introduction

The notion of derivatives in regular expressions is well established but has gained renewed attention in the last decade, for example in the work of Owens [8] and Might [5]. Their simplicity and compatibility with functional programming have encouraged further studies, for example the work of Sulzmann and Lu [9]. However, derivatives suffer from "growth" issues, since each step of taking the derivative can increase the size of subexpressions. This means the algorithm has to traverse larger and larger regular expressions, which results in a slower algorithm.

In contrast, the approach based on marks leaves regular expressions unchanged during matching but moves annotations through them (the marks). At present, only matcher-based algorithms using this approach exist, while our work aims to extend them to provide POSIX value extraction.

This report first reviews Brzozowski's derivatives and the bitcoded variant of Sulzmann and Lu [9], followed by background on the regular expression matchers based on marks described by Fischer et al. [4] and Asperti et al. [2]. We then present our own work, including several versions of the matcher based on marks developed during the first nine months.

2 Background

Regular expressions are a way to describe languages of words over an alphabet. They provide a declarative way to specify sets of words. Each regular expression r is associated with a language $L(r)$, which is the set of words that match r.

The regular expressions we consider are the basic ones with bounded repetition:

$$
\begin{array}{rcl}
r & ::= & \mathbf{0} \\
 & | & \mathbf{1} \\
 & | & c \quad \text{(character } c \text{ from the alphabet)} \\
 & | & r_1 + r_2 \\
 & | & r_1 \cdot r_2 \\
 & | & r^* \\
 & | & r^n
\end{array}
$$

We will use $\mathbf{0}$ for the regular expression of the empty language, $\mathbf{1}$ for the regular expression of the empty string, and c for the regular expression of

a character from the alphabet. The operator $+$ is for alternation between r_1 and r_2, while the dot $\cdot$ is for concatenation. The Kleene star r^* is for arbitrary repetition of r, and the power r^n is for exactly n repetitions of r.

The meaning of these regular expressions is given by the language function L, defined as:

Definition 1 (Language function L).

$$
\begin{aligned}
L(\mathbf{0}) &\overset{def}{=} \mathbf{0} \\
L(\mathbf{1}) &\overset{def}{=} \{[]\} \\
L(c) &\overset{def}{=} \{[c]\} \\
L(r_1 + r_2) &\overset{def}{=} L(r_1) \cup L(r_2) \\
L(r_1 \cdot r_2) &\overset{def}{=} L(r_1) \cdot L(r_2) \\
L(r^*) &\overset{def}{=} (L(r))^* \\
L(r^n) &\overset{def}{=} (L(r))^n
\end{aligned}
$$

For example, if $r = a + b$, then

$$L(r) = L(a) \cup L(b) = \{a\} \cup \{b\} = \{a, b\}.$$

Similarly, $L(a \cdot b) = \{ab\}$, and $L(a^*)$ is the set of all strings consisting of zero or more a's, including the empty string.

2.1 Derivatives

Brzozowski's derivatives offer an elegant way for string matching. By successively taking the derivative of a regular expression with respect to each input character, one obtains a sequence of derivatives. If the final expression can match the empty string, then the original input is accepted.

To decide whether a string $a_1 a_2 \ldots a_n$ is in the language of a regular expression r, we successively compute derivatives:

$$r_0 = r, \quad r_1 = der_{a_1}(r_0), \quad \ldots, r_n = der_{a_n}(r_{n-1}).$$

The string matches r if and only if the final expression r_n accepts the empty string. Here $der_a(r)$ stands for the derivative of r with respect to the character a, as introduced by Brzozowski [3] and defined below.

Definition 2 (Brzozowski's Derivative).

$$der_a(\boldsymbol{0}) \stackrel{def}{=} \boldsymbol{0}$$

$$der_a(\boldsymbol{1}) \stackrel{def}{=} \boldsymbol{0}$$

$$der_a(c) \stackrel{def}{=} \begin{cases} \boldsymbol{1} & if\ a = c \\ \boldsymbol{0} & if\ a \neq c \end{cases}$$

$$der_a(r_1 + r_2) \stackrel{def}{=} der_a(r_1) + der_a(r_2)$$

$$der_a(r_1 \cdot r_2) \stackrel{def}{=} der_a(r_1) \cdot r_2 + \begin{cases} der_a(r_2) & if\ \boldsymbol{1} \in L(r_1) \\ \boldsymbol{0} & otherwise \end{cases}$$

$$der_a(r^*) \stackrel{def}{=} der_a(r) \cdot r^*$$

To illustrate how derivatives can be used to match a regular expression against a string, consider the regular expression $(ab+ba)$. The derivative can check the matching of the string `ba` by taking the derivative of the regular expression with respect to b, then a.

$$\begin{aligned} der_b\, r &= der_b\, (ab + ba) \\ &= der_b\, (ab) + der_b\, (ba) \\ &= (der_b\, a) \cdot b + (der_b\, b) \cdot a \\ &= \boldsymbol{0} \cdot b + \boldsymbol{1} \cdot a \end{aligned}$$

$$\begin{aligned} der_a\, (der_b\, r) &= der_a\, (\boldsymbol{0} \cdot b + \boldsymbol{1} \cdot a) \\ &= der_a\, (\boldsymbol{0} \cdot b) + der_a\, (\boldsymbol{1} \cdot a) \\ &= der_a\, (\boldsymbol{0}) \cdot b + (der_a\, (\boldsymbol{1}) \cdot a + der_a\, a) \\ &= \boldsymbol{0} \cdot b + (\boldsymbol{0} \cdot a + \boldsymbol{1}) \end{aligned}$$

Since the final derivative expression contains $\boldsymbol{1}$, it matches the empty string. Because no input characters remain, this confirms that the original string `ba` is in the language of the regular expression.

Some subexpressions that arise during the computation of derivatives may be redundant. For example, $\boldsymbol{0} \cdot b$ expresses matching the empty language followed by b. Since $\boldsymbol{0}$ is the empty language, no string can satisfy this, so $\boldsymbol{0} \cdot b$ simplifies to $\boldsymbol{0}$.

Other simplification rules may also be applied. These include associativity $(r+s)+t \equiv r+(s+t)$, commutativity $r+s \equiv s+r$, and idempotence $r+r \equiv r$. Rules for the empty language and the empty string can also be applied, such as those mentioned earlier; e.g. $\boldsymbol{0} \cdot r \equiv r \cdot \boldsymbol{0} \equiv \boldsymbol{0}$, $r + \boldsymbol{0} \equiv r$, and $r \cdot \boldsymbol{1} \equiv \boldsymbol{1} \cdot r \equiv r$.

These simplifications preserve the language accepted by the regular expression while reducing the number of intermediate expressions. Such simplifications are sometimes necessary in the derivative method in order to keep the size manageable.

2.1.1 Size Explosion.

Although the construction of derivatives is elegant, as can be seen, it may duplicate large parts of the expression as well as the sizes of the regular expressions in the intermediate steps, and it is highly sensitive to the syntactic form of the expression and to nesting. Sulzmann and Lu [9] also note this problem, referring to it as size explosion. The size of derivatives can become very large even for simple expressions. This happens because each derivative is usually more complicated than the original expression, and even when basic simplifications are applied, derivatives may reintroduce the same sublanguage in different syntactic forms (e.g. $a + aa$ versus $a \cdot (1 + a)$), which simple rules and duplication removal do not recognise as equal. Consider the case of the regular expression $(a + aa)^*$. Each derivative step may introduce new structure, unfolding all possible ways in which the r^* expression can match the input. Even with basic simplifications, many of the resulting subexpressions still describe the same language but in different syntactic forms, and so the size continues to grow.

It is already known that simplification reduces the number of generated expressions and helps to provide a finite bound on the number of intermediate expressions [9, 10], but it does not completely solve the underlying problem. Even with aggressive simplifications—as in Sulzmann and Lu's bitcoded approach and in the variant described by Tan and Urban—the resulting derivatives can still be structurally large even if finitely bounded. For example, expressions of the form

$$((a)^* + (aa)^* + (aaa)^* + (aaaa)^* + \ldots)^*,$$

cannot be collapsed by simplification, since the duplicates occur in different syntactic forms. Thus, even under aggressive simplification the size of derivatives remains large and continues to grow across derivative steps.

Even Antimirov's partial derivatives [1], which do not track POSIX values, may still produce cubic growth in the worst case: there can be linearly many distinct derivatives, and each one may already be quadratic in the size of the original expression, so the total size becomes cubic.

7

2.1.2 Derivative Extension

Sulzmann and Lu [9] extend Brzozowski's derivatives to produce lexing values in addition to deciding whether a match exists. These values record how the match occurred, for example by showing which part of the regular expression matched the input string, whether the left or right branch was taken, and how sequences and Kleene stars were matched.

Sulzmann and Lu provide two variants: a bitcode-based construction and an injection-based construction [9]. In the bitcode variant, bit sequences encode lexing choices during derivative construction and, after acceptance, are decoded to the value. In the injection variant, an *inj* function "injects back" the consumed characters into the value; the injection function essentially reverts the derivative steps to obtain a POSIX value. We focus here on the bitcode variant, which more directly inspired our marked approach.

Bitcodes are sequences over $\{0, 1\}$ that encode the choices made in a match. These bits record branch choices in alternatives and repetitions made during matching. The following example illustrates how bitcoding works with regular expression $(a + ab)(b + ONE)$ and string ab. We proceed by constructing bitcoded derivatives step by step and tracking how the bitcode grows.

Initially, the algorithm internalizes the regular expression. Internalization only adds bit annotations to the alternative constructors found in the expression, where the left branch is annotated with bitcode 0 and the right branch with bitcode 1 [9]. This annotation process is implemented by the function *fuse*, whose definition is given at the end of this section. After internalization, the algorithm proceeds by taking the derivative with respect to the first character of the input string, which in this case is a.

1. Step 1: Internalizing the regular expression

$$r' \quad = \quad intern(r) = (_0a +\ _1ab) \cdot (_0b +\ _1\mathbf{1})$$

2. Step 2: input a

$$
\begin{aligned}
der_a(r') \quad &= \quad der_a((_0a +\ _1ab) \cdot (_0b +\ _1\mathbf{1})) \\
&= \quad der_a(_0a +\ _1ab) \cdot (_0b +\ _1\mathbf{1}) \\
&= \quad (der_a(_0a) + der_a(_1ab)) \cdot (_0b +\ _1\mathbf{1}) \\
&= \quad (_0\mathbf{1} + der_a(_1a) \cdot b) \cdot (_0b +\ _1\mathbf{1}) \\
&\rightarrow \quad (_0\mathbf{1} +\ _1\mathbf{1} \cdot b) \cdot (_0b +\ _1\mathbf{1})
\end{aligned}
$$

3. Step 3: input b

$$
\begin{aligned}
der_b(der_a(r')) &= der_b(({}_0\mathbf{1} + {}_1\mathbf{1} \cdot b) \cdot ({}_0 b + {}_1\mathbf{1})) \\
&= der_b({}_0\mathbf{1} + {}_1\mathbf{1} \cdot b) \cdot ({}_0 b + {}_1\mathbf{1}) + der_b(fuse(mkeps(r_1), ({}_0 b + {}_1\mathbf{1}))) \\
&= (der_b({}_0\mathbf{1}) + der_b({}_1\mathbf{1} \cdot b)) \cdot ({}_0 b + {}_1\mathbf{1}) + {}_0(der_b({}_0 b) + der_b({}_1\mathbf{1})) \\
&= (\mathbf{0} + der_b({}_1\mathbf{1} \cdot b)) \cdot ({}_0 b + {}_1\mathbf{1}) + {}_0({}_0\mathbf{1} + \mathbf{0}) \\
&= (der_b({}_1\mathbf{1}) \cdot b + der_b({}_1 b)) \cdot ({}_0 b + {}_1\mathbf{1}) + {}_0({}_0\mathbf{1} + \mathbf{0}) \\
&\rightarrow (\mathbf{0} \cdot b + {}_1\mathbf{1}) \cdot ({}_0 b + {}_1\mathbf{1}) + {}_0({}_0\mathbf{1} + \mathbf{0})
\end{aligned}
$$

The result of Step 2 shows that the ONE symbols indicate a successful match, while the bitcode records how that match was obtained. One match is obtained by taking the left branch to match the string a, reflected in the bitcode $[0]$. The other match arises by taking the right branch, which matches the regular expression (ab). The function $mkeps$, as defined by Sulzmann and Lu [9], extracts the bitcode from nullable derivatives. A subexpression is $nullable$ if it can match the empty string. Its definition is given at the end of this section.

In Step 3, the resulting derivative expands into an alternative. This occurs because the concatenation has become nullable, which means the first component, r_1, may be skipped while matching. To account for this, the derivative expands into an alternative: the left branch assumes r_1 is not skipped, while the right branch assumes it is skipped and instead takes the derivative of r_1 directly. This illustrates why derivatives tend to grow in size. Sulzmann and Lu [9] use $fuse(mkeps(r_1))$ to include the bit annotations needed when r_1 is skipped; these bits, extracted by $mkeps$, indicate how r_1 matched the empty string.

After taking the derivative with respect to the entire input string, the algorithm checks whether the result is nullable. If so, it calls $mkeps$ to extract the bitcode indicating how the match was obtained. In this example, there are two possible matches, but the algorithm prefers the left one. Consequently, $mkeps$ returns the bitcode $[1, 1]$, which encodes the choice of the right branch in the first alternative of r_1 (matching the string ab), followed by the right branch in the second part of the concatenation (matching the empty string).

The final bitcode after calling $mkeps$ is $[1, 1]$. Decoding this against the original expression yields the POSIX value:

$$
Seq(Right(Seq(a, b)), Right(Empty))
$$

As mentioned earlier, the formal definitions of the auxiliary functions *intern, fuse,* and *mkeps* are given below, as defined by Sulzmann and Lu [9].

Definition 3 (*intern* function).

$$intern(\mathbf{0}) \ \stackrel{def}{=} \ \mathbf{0}$$

$$intern(\mathbf{1}) \ \stackrel{def}{=} \ \mathbf{1}$$

$$intern(c) \ \stackrel{def}{=} \ c$$

$$intern(r_1 + r_2) \ \stackrel{def}{=} \ fuse(\,[0],\ intern(r_1)) \ + \ fuse(\,[1],\ intern(r_2))$$

$$intern(r_1 \cdot r_2) \ \stackrel{def}{=} \ intern(r_1) \cdot intern(r_2)$$

$$intern(r^*) \ \stackrel{def}{=} \ (intern(r))^*$$

Definition 4 (*fuse* function).

$$fuse \ bs \ (c_{\,bs'}) \ \stackrel{def}{=} \ c_{\,(bs \cup bs')}$$

$$fuse \ bs \ ((r_1 + r_2)_{\,bs'}) \ \stackrel{def}{=} \ (r_1 + r_2)_{\,(bs \cup bs')}$$

$$fuse \ bs \ ((r_1 \cdot r_2)_{\,bs'}) \ \stackrel{def}{=} \ (r_1 \cdot r_2)_{\,(bs \cup bs')}$$

$$fuse \ bs \ (r^*_{\,bs'}) \ \stackrel{def}{=} \ (r^*)_{\,(bs \cup bs')}$$

$$fuse \ bs \ (r^n_{\,bs'}) \ \stackrel{def}{=} \ (r^n)_{\,(bs \cup bs')}$$

Definition 5 (*mkeps* function).

$$mkeps(\mathbf{1}_{\,bs}) \ \stackrel{def}{=} \ bs$$

$$mkeps((r_1 + r_2)_{\,bs}) \ \stackrel{def}{=} \ \begin{cases} bs \cup mkeps(r_1) & if \ nullable(r_1) \\[2ex] bs \cup mkeps(r_2) & otherwise \end{cases}$$

$$mkeps((r_1 \cdot r_2)_{\,bs}) \ \stackrel{def}{=} \ bs \cup mkeps(r_1) \cup mkeps(r_2)$$

$$mkeps((r^*)_{\,bs}) \ \stackrel{def}{=} \ bs \cup [1]$$

2.2 Marked Approach

The marked approach is a method for regular expression matching that tracks matching progress by inserting and moving marks into the expression. As noted by Nipkow and Traytel [6], the idea can be traced to earlier work, in particular to Yamada and Glushkov, who identify positions in regular expressions by marking their atoms. Nipkow and Traytel cite Fischer et al. [4] and Asperti et al. [2] as reviving this work and developing it in a modern setting.

This approach allows for more efficient matching, particularly in complex expressions such as $(a^*b^*)^*$, in large alternations like $(a + b + c + \cdots + z)^*$, or in nested choices such as $((a)^* + (aa)^* + (aaa)^* + (aaaa)^* + (aaaaa)^*)^*$. The regular expression itself does not grow in size; instead, marks are inserted into it. With each input character, these marks move (or shift) according to a set of rules. At the end of the input, the expression is evaluated to determine whether the marks are in positions that make the expression "final" — that is, whether the expression accepts the string.

There is a slight difference in how marks are interpreted in the works of Fischer et al. [4] and Asperti et al. [2]. In Fischer et al.'s approach, marks are inserted after a character has been matched, thereby recording the matched character or subexpression. In contrast, Asperti et al. interpret the positions of marks as indicating the potential to match a character: as input is consumed, the marks move through the regular expression to indicate the next character that can be matched. In both approaches, acceptance is determined by evaluating the final state of the regular expression. For Fischer et al., this corresponds to having matched characters, whereas for Asperti et al. it requires that the marks end in positions where the expression can accept the empty string, ensuring that the entire input has been consumed. If the marks do not reach such positions, some characters remain unmatched and the matching problem fails.

2.2.1 Motivation for a Marked Approach

The marked approach offers an alternative to derivatives for regular expression matching. It relies on propagating marks within the regular expression rather than constructing new subexpressions. Our main motivation is that this method could support fast and high-performance matching with POSIX value extraction, since it handles matching in a way that avoids some of the

limitations of the derivative-based approach.

In the derivative method—for example, in the sequence case—the size of the expression typically increases due to the creation of new subexpressions, which contributes to the size explosion problem. By contrast, in the marked approach, matching achieves a similar result by propagating marks through the regular expression without generating larger expressions. We also hope that these marks can be used to extract POSIX values in an efficient manner.

Inspired by the works of Fischer et al. and Asperti et al. [4, 2], we aim to extend the marked approach to extract POSIX values, as well as to handle extended constructors, such as bounded repetitions and intersections. Our work is primarily based on the algorithm described by Fischer et al. [4]. As shown by Nipkow and Traytel [6], the pre-mark algorithm of Asperti et al. is in fact a special case of the post-mark algorithm of Fischer et al., which makes Fischer et al.'s approach the most suitable foundation for extending the marked algorithm to matching with POSIX value extraction.

2.2.2 Scala Implementation of Fischer's Marked Approach

In Fischer et al. approach, the marks are shifted through the regular expression with each input character. The process starts with an initial mark inserted at the beginning, which is then moved step by step as the input is consumed. This behaviour is implemented by the function *shift*, which performs the core logic of the algorithm. The initial specification of this function is given below, as we have developed several versions throughout our work.

The following presents the Scala implementation of the shifting behaviour as originally defined by Fischer et al. [4]. The *shift* function takes as input a regular expression to match against, a flag m, and a character c, and returns a *marked regular expression*—that is, a regular expression annotated with marks. We write a marked regular expression as $\bullet, r$, where the preceding dot indicates that the expression r has been annotated with marks to record the progress of matching.

$$shift(m, c, r) = \begin{cases} \bullet\, r & \text{if } c \text{ matches in } r, \\ \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

The flag m indicates the mode of operation: when set to true, a new mark is introduced; otherwise, the function shifts the existing marks. This

was realised in our first implementation by adding a boolean attribute to the character constructor to represent a marked character. In later versions, we instead introduced a wrapper constructor around the character constructor to explicitly represent a marked character.

Definition 6 (*shift* function).

$$shift(m, c, \mathbf{0}) \quad \overset{def}{=} \quad \mathbf{0}$$

$$shift(m, c, \mathbf{1}) \quad \overset{def}{=} \quad \mathbf{1}$$

$$shift(m, c, d) \quad \overset{def}{=} \quad \begin{cases} \bullet d & \text{if } c = d \wedge m \\ \\ d & \text{otherwise} \end{cases}$$

$$shift(m, c, r_1 + r_2) \quad \overset{def}{=} \quad shift(m, c, r_1) + shift(m, c, r_2)$$

$$shift(m, c, r_1 \cdot r_2) \quad \overset{def}{=}$$

$$\begin{cases} shift(m, c, r_1) \cdot shift(true, c, r_2) & \text{if } m \wedge nullable(r_1) \\ \\ shift(m, c, r_1) \cdot shift(true, c, r_2) & \text{if } fin(r_1) \\ \\ shift(m, c, r_1) \cdot shift(false, c, r_2) & \text{otherwise} \end{cases}$$

$$shift(m, c, r^*) \quad \overset{def}{=} \quad \begin{cases} shift(true, c, r^*) & \text{if } fin(r) \\ \\ shift(m, c, r^*) \end{cases}$$

Shifting marks for the base cases $\mathbf{0}$ and $\mathbf{1}$ is straightforward: $\mathbf{0}$ cannot be marked, and $\mathbf{1}$ —for now—will not carry a mark, since it matches only the empty string. In the initial algorithm, $\mathbf{1}$ was not marked, and this choice is carried over into later versions. The reason is to avoid complications and ensure termination of the *shift* function, some made apparent in the latter versions as we will discuss further in subsequent sections.[1] The behaviour of the remaining cases is described next.

---

[1] Our choice follows Fischer et al. [4], where $\mathbf{1}$ is left unmarked ('shift EPS = EPS'). Asperti et al. [2], on the other hand, use pointed regular expressions (pREs) where acceptance of the empty string is represented by the trailing point being set to true.

- **Character case** (d): In this case, if the input character c matches d and the flag m is true, a mark is added and stored in the character constructor. Otherwise, the character remains unmarked.

- **Alternative case** $(r_1 + r_2)$: Marks are shifted into both subexpressions, since either branch may match the input character.

- **Sequence case** $(r_1 \cdot r_2)$:
  - If r_1 is neither nullable nor in a final position, marks are shifted only into r_1, indicating that matching proceeds with the first component.
  - If r_1 is nullable and may be skipped, marks are shifted into both r_1 and r_2, so that either component can begin matching.
  - If $\mathit{fin}(r_1)$ holds, meaning r_1 has finished matching, marks are shifted into r_2 to continue matching with its component.

- **Star case** (r^*): Marks are shifted into the subexpression if m is true or if $\mathit{fin}(r)$ holds.

The formal definitions of the auxiliary functions fin and $\mathit{nullable}$ are given below, as defined by Fischer et al. [4].

Definition 7 (fin function).

$$
\begin{aligned}
\mathit{fin}(\mathbf{0}) &\stackrel{def}{=} \mathit{false} \\
\mathit{fin}(\mathbf{1}) &\stackrel{def}{=} \mathit{false} \\
\mathit{fin}(c) &\stackrel{def}{=} \mathit{false} \\
\mathit{fin}(\bullet\, c) &\stackrel{def}{=} \mathit{true} \\
\mathit{fin}(r_1 + r_2) &\stackrel{def}{=} \mathit{fin}(r_1) \vee \mathit{fin}(r_2) \\
\mathit{fin}(r_1 \cdot r_2) &\stackrel{def}{=} (\mathit{fin}(r_1) \wedge \mathit{nullable}(r_2)) \vee \mathit{fin}(r_2) \\
\mathit{fin}(r^*) &\stackrel{def}{=} \mathit{fin}(r)
\end{aligned}
$$

Definition 8 ($\mathit{nullable}$ function).

$$
\begin{aligned}
nullable(\mathbf{0}) &\stackrel{def}{=} false \\
nullable(\mathbf{1}) &\stackrel{def}{=} true \\
nullable(c) &\stackrel{def}{=} false \\
nullable(r_1 + r_2) &\stackrel{def}{=} nullable(r_1) \vee nullable(r_2) \\
nullable(r_1 \cdot r_2) &\stackrel{def}{=} nullable(r_1) \wedge nullable(r_2) \\
nullable(r^*) &\stackrel{def}{=} true
\end{aligned}
$$

3   Our Approach

We began our work by implementing the marked approach described by Fischer et al. [4] in Scala. This initial implementation of the algorithm provided only acceptance checking without any value construction.

Over the course of this work, we developed several versions of the algorithm, each addressing specific challenges. The first version extended the marked approach with bit annotations, producing values but not always the POSIX values. This happens in cases where the marks gets overwritten when a character is matched more than once.

The second version was developed to address two main issues we faced: the overwriting of marks and the absence of a mechanism to order them so as to preserve the POSIX value. Initially, we modified the previous version to accumulate all possible paths to a match, retaining every bit sequence that could lead to acceptance. Through reasoning and testing, we kept refining it to the point where we are fairly confident that it can produce all possible values for a given string and regular expression, including the POSIX value. One resulting difficulty was the proliferation of matches in the *Star* (r^*) case, where every possible way of matching was generated. Another difficulty, which also existed in the first version, was the absence of an ordering for marks during the shifting process. We implemented an ordering after shifting to evaluate the results of the algorithm, following the work of Okui and Suzuki [7], though not inherently in the shifting process itself. This line of thought eventually led us to the final version of the algorithm, which uses string-annotated marks, where marks carry the matched string along as they are shifted through the regular expression and then added bit annotation

to the marks to record the choices made during matching. The following subsections describe these different versions.

3.1   Bit-Annotated, version 1

In this version, we extended the marked approach to include bitcodes that annotate the marks being shifted through the regular expression. Inspired by Sulzmann and Lu [9], we introduced bitcodes in the form of lists attached to each mark, which are incrementally built as the marks are shifted. This version produces a value, though not necessarily the POSIX value, because when a character is matched more than once at the same point, the associated bit list may be overwritten. This can cause value erasure and, in some cases, the loss of the POSIX value, as illustrated later in Example 2. We use the bit annotations 0 and 1, similar to the bitcoded derivatives described by Sulzmann and Lu [9].

The function *shift* takes an additional argument, *Bits*, which is a list of bit elements (0 or 1). As *shift* is applied, bits are appended to the list. For example, when shifting through an alternative, 0 is added to the list passed to the left subexpression and 1 to the list passed to the right subexpression. If the input character matches a leaf character node, it is wrapped by the newly defined *POINT* constructor, which represents the mark. The associated bit list is stored inside this constructor together with the character.

We define the auxiliary function *mkfin* to extract the bit sequence of a final constructor – that is, the path in bits describing how the expression matched. We adapt *mkeps* from Sulzmann and Lu [9] and from Tan and Urban [10], who define it to construct a value tree and a bit sequence, respectively, for how a nullable expression matches the empty string. Our version also returns a bit sequence: 0 for the left branch and 1 for the right branch in choices, and for *Star* (r^*), 0 indicates the start of an iteration and 1 its end, with the single bit 1 used for the empty-star case. The definitions of *mkfin* and *mkeps* are given after the definition of *shift*.

We define *shift* for this version as follows.

$$
shift(m, c, bs, r) = \begin{cases} \bullet\, r_{\ bs'} & \text{if } c \text{ matches in } r, \\[2ex] \mathbf{0} & \text{otherwise.} \end{cases}
$$

where bs' is the updated bit list.

Definition 9 (Bit-Annotated *shift*: First Version)**.**

$$shift(m, bs, c, \mathbf{0}) \quad \overset{def}{=} \quad \mathbf{0}$$

$$shift(m, bs, c, \mathbf{1}) \quad \overset{def}{=} \quad \mathbf{1}$$

$$shift(m, bs, c, d) \quad \overset{def}{=} \quad \begin{cases} \bullet\, d_{bs} & \text{if } m \wedge d = c \\ d & \text{otherwise} \end{cases}$$

$$shift(m, bs, c, r_1 + r_2) \quad \overset{def}{=} \quad shift(m,\, bs \oplus 0,\, c,\, r_1) \;+\; shift(m,\, bs \oplus 1,\, c,\, r_2)$$

$$shift(m, bs, c, r_1 \cdot r_2) \quad \overset{def}{=}$$

$$\begin{cases} shift(m, bs, c, r_1) \cdot shift(true,\, bs\,@\,mkeps(r_1),\, c,\, r_2) & \text{if } m \wedge nullable(r_1) \\[1.5em] shift(m, bs, c, r_1) \cdot shift(true,\, bs\,@\,mkfin(r_1),\, c,\, r_2) & \text{if } fin(r_1) \\[1.5em] shift(m, bs, c, r_1) \cdot shift(false,\, [\,],\, c,\, r_2) & \text{otherwise} \end{cases}$$

$$shift(m, bs, c, r^*) \quad \overset{def}{=}$$

$$\begin{cases} (shift(m,\, bs \oplus 0,\, c,\, r))^* & \text{if } m \\[1.5em] (shift(true,\, bs\,@\,(mkfin(r) \oplus 1),\, c,\, r))^* & \text{if } m \wedge fin(r) \\[1.5em] (shift(true,\, mkfin(r) \oplus 0,\, c,\, r))^* & \text{if } fin(r) \\[1.5em] (shift(false,\, [\,],\, c,\, r))^* & \text{otherwise} \end{cases}$$

Here, $\oplus$ stands for appending a bit to a list, while @ concatenates two lists. The symbols 0 and 1 are used to represent left and right choices in alternatives. In the *Star* (r^*) case, 0 represents the beginning of an iteration, while 1 represents the end of an iteration.

Alternative case $(r_1 + r_2)$:

- Marks are shifted as before, and the direction of the match is annotated: 0 is added to the bit list passed to the left subexpression, and 1 is added to the bit list passed to the right subexpression.

Sequence case $(r_1 \cdot r_2)$**:**

- If r_1 is nullable, a mark is shifted to both r_1 and r_2: bs is passed to r_1 (representing the path to this expression), and $bs \cup mkeps(r_1)$ is passed to r_2, where *mkeps* returns the bits for an empty-string match. This corresponds to the case where the first part of the sequence is skipped.

- If r_1 is in a final position (that is, it has finished matching), a mark is shifted to r_2 with the bit list describing how r_1 was matched, extracted using the *mkfin* function.

- Otherwise, marks are shifted only into r_1, with bs representing the current path to r_1.

Star case (r^*)**:**

- If a new mark is introduced, bs is passed with 0 appended, representing the beginning of a new iteration of the star.

- If a new mark is introduced and r is in a final position, $bs \cup mkfin(r)$ is passed with 1 appended, combining the bits describing the path to r^* with the bits showing how r reached a final position.

- If r is in a final position, $mkfin(r)$ is passed with 0 appended, representing the start of a new iteration.

- If no new mark is introduced, the existing marks are shifted with an empty bit list.

The functions *mkfin* and *mkeps* are defined below; the definition of *fin* is the same as in Section 2.2.2.

Definition 10 (*mkfin* function)**.**

$$mkfin(\bullet_{bs}\, r) \quad \overset{def}{=} \quad bs$$

$$mkfin(\bullet_{bs}\, c) \quad \overset{def}{=} \quad bs$$

$$mkfin(r_1 + r_2) \quad \overset{def}{=} \quad \begin{cases} mkfin(r_1) & \text{if } fin(r_1) \\ \\ mkfin(r_2) & otherwise \end{cases}$$

$$mkfin(r_1 \cdot r_2) \quad \overset{def}{=} \quad \begin{cases} mkfin(r_1)\, @\, mkeps(r_2) & \text{if } fin(r_1) \wedge nullable(r_2) \\ \\ mkfin(r_2) & else \end{cases}$$

$$mkfin(r^*) \quad \overset{def}{=} \quad mkfin(r)\, @\, [1]$$

Definition 11 (*mkeps* function)**.**

$$mkeps(\bullet_{bs}\, r) \quad \overset{def}{=} \quad bs$$

$$mkeps(\mathbf{1}) \quad \overset{def}{=} \quad [\,]$$

$$mkeps(r_1 + r_2) \quad \overset{def}{=} \quad \begin{cases} 0 \oplus mkeps(r_1) & \text{if } nullable(r_1) \\ \\ 1 \oplus mkeps(r_2) & else \end{cases}$$

$$mkeps(r_1 \cdot r_2) \quad \overset{def}{=} \quad mkeps(r_1)\, @\, mkeps(r_2)$$

$$mkeps(r^*) \quad \overset{def}{=} \quad [1]$$

- $mkfin(\bullet_{bs}\, r) \to bs$

$$mkfin(\bullet_{bs}\, c) \quad \overset{\text{def}}{=} \quad bs$$

$$mkfin(r_1 + r_2) \quad \overset{\text{def}}{=} \quad \begin{cases} mkfin(r_1) & \text{if } fin(r_1) \\ mkfin(r_2) & \text{otherwise} \end{cases}$$

$$mkfin(r_1 \cdot r_2) \quad \overset{\text{def}}{=} \quad \begin{cases} mkfin(r_1)\, @\, mkeps(r_2) & \text{if } fin(r_1) \wedge nullable(r_2) \\ mkfin(r_2) & \text{else} \end{cases}$$

$$mkfin(r^*) \quad \overset{\text{def}}{=} \quad mkfin(r)\, @\, [1]$$

- $mkeps(\bullet_{bs}\, r) \to bs$

$$
\begin{aligned}
mkeps(\\
ONE\,) &\stackrel{\text{def}}{=} [\,] \\
mkeps(r_1 + r_2) &\stackrel{\text{def}}{=}
\begin{cases} 0 \oplus mkeps(r_1) & \text{if } nullable(r_1) \\ 1 \oplus mkeps(r_2) & \text{else} \end{cases} \\
mkeps(r_1 \cdot r_2) &\stackrel{\text{def}}{=} mkeps(r_1) \,@\, mkeps(r_2) \\
mkeps(r^*) &\stackrel{\text{def}}{=} [1]
\end{aligned}
$$

Next, we present two examples of matching a string and extracting a value. The first example shows how the bit list is constructed during matching, while the second demonstrates a case where the algorithm fails to produce the POSIX value. When shifting to a point (an already marked character) and the character matches again, the associated bit list is overwritten which can lead to value erasure and, in particular, to the loss of the POSIX value, as the second example illustrates.

1. String ba, regular expression: $(a \cdot b + b \cdot a)$

$$shift\ b \quad \to \quad (a \cdot b) + (\bullet_{[1]}\, b \cdot a)$$

$$shift\ a \quad \to \quad (a \cdot b) + (b \cdot \bullet_{[1]}\, a)$$

With no further calls to *shift*, *mkfin* is applied because the regular expression has reached a final position, indicated by a mark at the end of the right-hand subexpression in the alternative. *mkfin* then retrieves the bit list [1], corresponding to taking the right branch in the alternative.

2. String aaa, regular expression: $(a + a \cdot a)^*$

$$shift\ a \quad \to \quad (\bullet_{[0,0]}\, a + \bullet_{[0,1]}\, a \cdot a)^*$$

$$shift\ a \quad \to \quad (\bullet_{[0,0,0,0]}\, a + \bullet_{[0,0,0,1]}\, a \cdot \bullet_{[0,1]}\, a)^*$$

$$shift\ a \quad \to \quad (\bullet_{[0,0,0,0,0,0]}\, a + \bullet_{[0,0,0,0,0,1]}\, a \cdot \bullet_{[0,0,0,1]}\, a)^*$$

After the first shift on a, a mark is placed on the left branch with bits $[0,0]$, indicating the start of a *Star* (r^*) iteration followed by a left

choice. In the right subexpression, the mark on r_1 of the $a \cdot a$ sequence carries bits $[0, 1]$, representing the start of the *Star* (r^*) iteration followed by a right choice.

After the second shift, however, the right subexpression r_2 with bits $[0, 1]$ is overwritten during the third shift, when those bits should instead be preserved. These bits correspond to the POSIX match, which starts by matching the right-hand side first, then performing another iteration to match the left-hand side. The correct bit list in that case would be $[0, 1, 0, 0]$, with the final 1 marking the end of the *Star* iteration. This behaviour arises because $POINT$ stores only a single bit list at a time, with no mechanism for preserving multiple marks.

3.2 Bit-Annotated, version 2

we are fairly sure/strongly think that this version produces all possible values including the posix value.

3.3 String-Carrying Marks - Matcher

We initially followed the approach of Fischer et al. [4] and Asperti et al. [2], in which shifting is performed character by character and marks are propagated at each step. However, as in our previous versions, this method made it difficult to maintain an ordering on marks and to extract the POSIX value.

To address this, we instead let each mark carry the suffix of the input string that is still to be matched, starting with an initial mark carrying the full input string. It gives a clear way of ordering the marks, based on the remaining suffixes. It also appears to offer a closer correspondence with derivatives, in that each mark could be associated with subexpressions created by the derivative, though this point is not yet fully established. The trade-off is that more information must be stored, but this also creates opportunities for optimisation. With string-carrying marks, we currently have a better handle on extracting POSIX values.

We start with a list of marks with one initial mark carrying the full input string. As *shift* is called, the marks are shifted through the regular expression, and each time a character matches, it is striped from the string carried by the mark. If a character does not match the prefix of the string in a mark, that mark will be dropped. A match occurs when a mark has been reduced to the empty string, indicating that all characters have been matched. When the input string itself is empty, *shift* is not applied, and instead the *nullable* function is used to check whether the regular expression accepts the empty string.

The number of marks in this version can grow quadratically in the worst case. For a regular expression of size n, we can have up to n marks at any given time in most cases. The exception is the *Star* (r^*) case, where each unfolding of the star may generate as many as n marks, so that with n unfoldings the total number of marks can reach n^2. In the worst case this leads to a quadratic bound once duplicate marks are removed.

We started with a lexer implementation of this version, and from initial testing it produces the POSIX value, yet this remains to be formally proven. In this version we introduce reshuffling, where marks are ordered based on

their remaining strings. At present, the reordering is carried out in the *Sequence* $(r_1 \cdot r_2)$ case after shifting through the first part.

We define *shift* for this version as follows; the behaviour of the individual cases is then described.

$$
\begin{aligned}
shift(ms, 0) &\stackrel{\text{def}}{=} [\,] \\
shift(ms, 1) &\stackrel{\text{def}}{=} [\,] \\
shift(ms, d) &\stackrel{\text{def}}{=} [\, \bullet s \mid \bullet d :: s \in ms \,] \\
shift(ms, r_1 + r_2) &\stackrel{\text{def}}{=} shift(ms, r_1) \,@\, shift(ms, r_2) \\
shift(ms, r_1 \cdot r_2) &\stackrel{\text{def}}{=} \text{let } ms' = shift(ms, r_1) \text{ in} \\
&\qquad
\begin{cases}
shift(ms'@ms, r_2) \,@\, ms' & \text{if } nullable(r_1) \wedge nullable(r_2), \\
shift(ms'@ms, r_2) & \text{if } nullable(r_1), \\
shift(ms', r_2) \,@\, ms' & \text{if } nullable(r_2), \\
shift(ms', r_2) & \text{otherwise}
\end{cases} \\
shift(ms, r^*) &\stackrel{\text{def}}{=} \text{let } ms' = shift(ms, r) \text{ in} \\
&\qquad \text{if } ms' = [\,] \text{ then } [\,] \text{ else } shift(ms', r^*) \,@\, ms'
\end{aligned}
$$

As noted earlier, the empty string **1** is not marked. This choice is made to ensure termination of the *shift* function, in particular in the *Star* (r^*) case: if ONE were allowed to produce marks, this would lead to infinite unfolding and non-termination. A Mark is represented here as $\bullet s$, where s is a string, and ms is a list of such marks.

Alternative case $(r_1 + r_2)$**:**

- The list of marks ms is shifted into both subexpressions r_1 and r_2. The two resulting lists are then concatenated, which corresponds to matching with either r_1 or r_2.

Sequence case $(r_1 \cdot r_2)$**:**

- This is the most elaborate case. It begins by shifting the marks into the first part of the sequence, r_1, producing a new list of marks ms'. The result then depends on several conditions:, namely:

23

1. *nullable*$(r_1 \wedge r_2)$: both r_1 and r_2 are nullable, so both can be skipped and the shifting must account for this by combining the results appropriately. To account for r_1 being skipped, the original marks ms are appended to ms' and then shifted into r_2. In this way, skipping r_1 has the same effect as shifting ms directly into r_2. Finally, ms' itself is appended to the result to account for skipping r_2, which corresponds to matching with r_1 only.

2. *nullable*(r_1): if r_1 is nullable, it may be skipped. In this case, the original marks ms are appended to ms' and the result is then shifted into r_2, which corresponds to matching directly with r_2.

3. *nullable*(r_2): if r_2 is nullable, it may be skipped. Here, ms' is shifted into r_2, and the result is then appended with ms' to account for the case where r_2 is skipped, which corresponds to matching with r_1 only.

4. $\neg$*nullable*$(r_1) \wedge \neg$*nullable*(r_2): if neither r_1 nor r_2 is nullable, the marks ms' are shifted directly into r_2.

Star case (r^*)**:**

- The list of marks ms is first shifted into r, producing a new list ms'. If ms' is empty, this corresponds to the end of the iterations of the *Star* (r^*), and the result is the empty list. Otherwise, ms' records one completed iteration and is appended to the result of shifting ms' into the *Star* (r^*), corresponding to adding a new iteration.

The matcher is defined as follows.

$$matcher(s, r) \stackrel{\text{def}}{=} \begin{cases} nullable(r) & \text{if } s = [\,], \\ \bullet [\,] \in shift([\bullet\, s], r) & \text{otherwise.} \end{cases}$$

To illustrate the behaviour of this version, we consider two examples. The first uses the regular expression $b + (b \cdot a)$ and string `ba`. We show step by step how the marks are reduced as characters are stripped and non-matching paths are dropped. The second example considers the star case, highlighting how the marks are unfolded.

1. String `ac`, regular expression: $(a + (a \cdot c))$

24

(a) $\big|_{[\bullet\, ac]} (a \,+\, (a \cdot c))$

(b) $\big|_{[\bullet\, ac]} a \,+\, \big|_{[\bullet\, ac]} a \cdot c$

(c) $a \,\big|_{[\bullet\, c]} + a \cdot c \,\big|_{[\bullet\, []]}$

Matching begins with the initial mark carrying the full string, which is then propagated into the two subexpressions. In the first subexpression, the initial character matches, so it is stripped from the mark, leaving the string c. In the second subexpression, the sequence $a \cdot c$ matches both characters, so the string in the mark is reduced to the empty string. At this point, the list of marks $[\bullet_{[c]}, \bullet_{[]}]$ is obtained, and with one mark reduced to the empty string, the expression is accepted.

2. String aaa, regular expression: $(a)^*$

3. String aaa, regular expression: $(a + aa)^*$

3.4 String-Carrying Marks - Lexer **so far**

4 Future Work

*** from previous report ***

This project focuses on implementing and validating a correct and efficient marked regular expression matcher under POSIX disambiguation. Several directions remain open and are planned for the next stages of the PhD:

- **POSIX Disambiguation for STAR.** While the current matcher correctly computes POSIX values for many expressions, disambiguation for nested or ambiguous STAR patterns is not yet complete. Ensuring that the correct POSIX value is selected in all cases involving repetition remains a primary target. The current implementation explores candidate paths, but the disambiguation logic for selecting among them requires refinement and formal confirmation.

- **Support for Additional Operators.** Beyond the basic constructs (ALT, SEQ, STAR, NTIMES), future work includes extending the matcher to handle additional regex operators such as intersection, negation, and lookahead. These additions require careful definition of how

marks behave and how disambiguation should be handled, but could significantly increase the expressiveness of the engine.

- **Formal Proof of POSIX Value Correctness.** A formal verification is planned to prove that the marked matcher always produces the correct POSIX-disambiguated value. This would involve defining the decoding function rigorously and proving its output corresponds to the POSIX parse. This direction is part of the original PhD proposal, where value extraction and correctness proofs were identified as key goals.

References

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

[2] A. Asperti, C. S. Coen, and E. Tassi. Regular Expressions, au point. *arXiv, http://arxiv.org/abs/1010.2604*, 2010.

[3] J. A. Brzozowski. Derivatives of regular expressions. *J. ACM*, 11(4):481–494, Oct. 1964.

[4] S. Fischer, F. Huch, and T. Wilke. A Play on Regular Expressions: Functional Pearl. In *Proc. of the 15th ACM SIGPLAN International Conference on Functional Programming (ICFP)*, pages 357–368, 2010.

[5] M. Might, D. Darais, and D. Spiewak. Parsing with derivatives: A functional pearl. In *Proceedings of the 16th ACM SIGPLAN International Conference on Functional Programming (ICFP)*, pages 189–195. ACM, 2011.

[6] T. Nipkow and D. Traytel. Unified decision procedures for regular expression equivalence. In G. Klein and R. Gamboa, editors, *Interactive Theorem Proving*, pages 450–466, Cham, 2014. Springer International Publishing.

[7] S. Okui and T. Suzuki. Disambiguation in regular expression matching via position automata with augmented transitions. Technical Report Technical Report 2013-002, The University of Aizu, Language Processing Systems Laboratory, June 2013.

[8] S. Owens, J. Reppy, and A. Turon. Regular-expression derivatives re-examined. *Journal of Functional Programming*, 19(2):173–190, 2009.

[9] M. Sulzmann and K. F. Lu. Posix regular expression parsing with derivatives. *Science of Computer Programming*, 89:179–193, 2014.

[10] C. Tan and C. Urban. *POSIX Lexing with Bitcoded Derivatives*, pages 26:1–26:18. Apr. 2023.