

Lost in Translation? Converting REGEXES for Log Parsing into Dynatrace Pattern Language

Julian Fragner^{a,*}, Christian Macho^b, Bernhard Dieber^a and Martin Pinzger^b

^a*Dynatrace Austria GmbH, Klagenfurt, Austria*

^b*University of Klagenfurt, Klagenfurt, Austria*

ARTICLE INFO

Keywords:

Log parsing
Regular expressions
Dynatrace Pattern Language
Log pattern conversion
Log pattern optimization

ABSTRACT

Log files provide valuable information for detecting and diagnosing problems in enterprise software applications and data centers. Several log analytics tools and platforms were developed to help filter and extract information from logs, typically using regular expressions (REGEXES). Recent commercial log analytics platforms provide domain-specific languages specifically designed for log parsing, such as Grok or the Dynatrace Pattern Language (DPL). However, users who want to migrate to these platforms must manually convert their REGEXES into the new pattern language, which is costly and error-prone.

In this work, we present REPTILE, which combines a rule-based approach for converting REGEXES into DPL patterns with a best-effort approach for cases where a full conversion is impossible. Furthermore, it integrates GPT-4 to optimize the obtained DPL patterns. The evaluation with 946 REGEXES collected from a large company shows that REPTILE safely converted 73.7% of them. The evaluation of REPTILE's pattern optimization with 23 real-world REGEXES showed an F1-score and MCC above 0.91. These results are promising and have ample practical implications for companies that migrate to a modern log analytics platform, such as Dynatrace.

1. Introduction

Enterprise software applications and data centers become increasingly complex, which in turn increases the effort of operating them [1]. Often, seamless execution or agreed performance of software systems must be guaranteed [2]. In large organizations, dedicated teams are assigned to monitor software execution and identify system events at runtime. Log files are seen as a primary source for problem diagnosis [3]. They provide valuable runtime information of applications, which can be used for maintenance, troubleshooting, anomaly and problem detection, failure prediction, root cause analysis, performance diagnosis, and security threat detection [1, 2, 4, 5].

To avoid downtimes and meet service-level agreements (SLAs) of business applications and data centers, it is important to analyze log files quickly and accurately [2, 4]. However, with the increasing size and number of software systems in an organization, the volume and variety of logs grow [1, 2]. This makes manual inspection of log files practically infeasible [6–8]. Therefore, commercial log analytics tools and platforms, such as *Splunk*¹, *ElasticSearch*²,

*Datadog*³, and *Dynatrace*⁴, emerged to help filter and extract information from logs [1].

As no standardized format and official logging guidelines exist [1, 3], log messages are mostly unstructured and vary widely in content and format. To extract information from such heterogeneous logs, most log analytics platforms require human involvement to define log formats using customized patterns [4]. Regular expressions (REGEXES) are commonly used for this task [1, 2, 9]. However, creating correct REGEXES can be challenging, time-consuming, and error-prone, and requires specialized skills and experience [2, 10–21]. Additionally, the REGEX syntax is often difficult to comprehend [14, 18, 20, 22], making pattern maintenance cumbersome and error-prone [1, 2].

To address these drawbacks, some providers utilize domain-specific languages (DSL) specifically designed for log parsing. For example, *ElasticSearch* [23] and *Datadog* [24] use *Grok*, which builds on top of REGEXES, while *Dynatrace* uses the proprietary *Dynatrace Pattern Language* (DPL) [25, 26]. These languages are generally more user-friendly than REGEXES and require less technical background from the user. For example, they provide predefined matchers for commonly occurring log data, such as timestamps and IP addresses, which enhances efficiency of pattern creation and reduces the risk of human error.

Figure 1 demonstrates these advantages by comparing a sample log pattern in REGEX, Grok, and DPL syntax respectively. Line 2 shows the REGEX that extracts the IP address, matches any text followed by at least one space, and finally extracts the response code from a log entry. Lines 5 and 8 show its Grok and DPL counterparts, respectively,

*Corresponding author

 julian.fragner@dynatrace.com (J. Fragner);

christian.macho@aau.at (C. Macho); bernhard.dieber@dynatrace.com (B. Dieber); martin.pinzger@aau.at (M. Pinzger)

 <https://mitschi.github.io/> (C. Macho);

<https://www.bernharddieber.com/> (B. Dieber); <https://pinzger.github.io/> (M. Pinzger)

ORCID(s): 0000-0001-8182-7277 (C. Macho); 0000-0002-0450-8990 (B. Dieber); 0000-0002-5536-3859 (M. Pinzger)

¹<https://www.splunk.com/>

²<https://www.elastic.co/elasticsearch>

³<https://www.datadoghq.com/>

⁴<https://www.dynatrace.com/>

Figure 1: REGEX example and corresponding Grok and DPL pattern examples.

```

1 // RegEx
2 (?<addr>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}).*\s+(?<rc>\d{3})
3
4 // Grok counterpart
5
6 // DPL counterpart
7 IPv4:addr LD SPACE+ INT:rc
    
```

which are easier to read than the REGEX as they are shorter and contain fewer special characters. The DPL additionally separates its individual matchers by spaces. However, the use of different pattern languages creates problems when users want to migrate from one log analytics provider to another. The manually created patterns must then be manually translated into the target language which is costly and error-prone.

In this work, we present REPTILE⁵, an approach that converts REGEXES into DPL patterns. We choose REGEXES as the source language, as it is widely used by different log analytics and monitoring solutions as mentioned above and DPL as the target language, because this work was done in collaboration with Dynatrace which is a leading company in providing software observability solutions [27]. REPTILE facilitates automatic generation of DPL patterns purely from REGEXES whenever possible. It also provides best-effort conversions and only involves the user in cases for which automatic translation is not possible. As a first step towards REPTILE, we investigate the usage of REGEXES in practice to understand how practitioners use them. The corresponding research question is:

RQ1 Which REGEX features are frequently used in practice?

We studied a real-world dataset from a large organization from the retail industry and extracted 946 different REGEXES. For each REGEX, we tallied the used REGEX features. We found that *Named capturing group*, *Greedy quantifier*, and *Literal character* were the top-3 most used features (↔ Section 3.2). These findings laid the foundation for developing REPTILE that we evaluated in our second research question:

RQ2 How effective is REPTILE in converting REGEXES into DPL patterns?

The evaluation shows that our approach increases the percentage of safely convertible REGEXES from initially less than 2% to more than 73% (↔ Section 5.1) in a real-world dataset. Moreover, 897,000 generated test cases show that our approach retains the original REGEX semantics. These results show that our approach supports users in safely converting the majority of REGEXES into DPL. Next, we investigate the usage of GPT-4 to further optimize the DPL

patterns by introducing high-level matchers. This leads to our third research question:

RQ3 What is the prediction accuracy of REPTILE's pattern optimization?

The results with GPT-4 and 23 real-world REGEXES show high prediction performance achieving an average F1-score and MCC above 0.91 across five high-level DPL matchers (↔ Section 5.2). This means that the majority of high-level matchers proposed by GPT-4 are correct.

In summary, our work makes the following contributions:

- A study of most frequently used REGEX features for log parsing in practice, analyzing more than 900 real-world REGEXES (↔ Section 3.2).
- Novel strategies to identify non-backtracking REGEXES, enabling conversion of greedy and lazy quantifiers, which are the most frequently used features unsupported in DPL (↔ Section 4.1).
- A two-phase approach for automatic conversion of REGEXES into DPL, comprising of a rule-based conversion (↔ Section 4.1) and an optimization with GPT-4 (↔ Section 4.3).

Our results have ample practical implications for companies that migrate or migrated to a modern log analytics platform, such as Dynatrace. They can use REPTILE to automatically and safely convert the majority (in our case 73%) of their REGEXES for log parsing to DPL patterns. Furthermore, they can use REPTILE to optimize the converted DPL patterns making them easier to understand and maintain with high precision and recall.

The remainder of the paper is organized as follows. First, Section 2 defines terms and concepts. Next, we study the most frequently used REGEX features in Section 3. Section 4 presents the converter prototype REPTILE. Section 5 assesses the proportion of safely convertible REGEXES and their correctness and evaluates the accuracy of high-level matcher prediction. Section 7 situates the paper with respect to related work. Section 8 discusses the main results, current limitations, and potential threats to the validity. Finally, Section 8 concludes the paper and envisions future work.

2. Background

This section outlines the key features of REGEXES (↔ Section 2.1) and DPL (↔ Section 2.2). We only present the most relevant features for this study. For more detailed explanations, we refer the reader to standard works on REGEXES [14, 28] and the public DPL documentation [25].

2.1. Regular Expressions (REGEXES)

This section is based on the work of Friedl [14] and provides details of the most relevant REGEX features. Note that in this work, the term REGEXES does not strictly refer to the theoretical concept of regular expressions (first

⁵Regular Expression Pattern Translation Into Language Equivalents

Table 1
Overview of greedy quantifiers available in REGEXES

Quantifier	Equivalent	Match
?	{0,1}	zero times or once
*	{0,}	zero, one, or multiple times
+	{1,}	once or multiple times
{x}	{x,x}	exactly x times
{x,y}	-	min. x, max. y times
{x,}	-	min. x times

introduced by Kleene [29] in 1956) as a way to express a regular language [30, 31]. Instead, it refers to the practical implementation of regular expressions that are commonly used in programming languages, standard libraries, and text-processing applications [17, 19, 20]. These practical implementations often support *non-regular* features that extend the capabilities of regular languages, such as backreferences and lookarounds [18, 32]. Our investigation is limited to the widespread PCRE (Perl-compatible regular expressions) flavor [14, 33]. A comprehensive comparison and implementation of different flavors is beyond the scope of this work.

2.1.1. Backtracking Quantifiers

Quantifiers are a central feature of REGEXES. They indicate that their preceding matcher can match optionally, or more than only once. See Table 1 for an overview of available quantifiers. Of all quantifiers, the most frequently used are the *greedy* ones, which match as much of the input as possible. An example of this is shown in Figure 2, where the match result may seem counterintuitive. The first line with gray background contains the REGEX. Below that is the input string to which the REGEX is applied, which we refer to as the *target string*. The match result is indicated by dashes below the target string. The set of all strings accepted by a given REGEX is referred to as its *defined language*. We continue to use these typographical and terminological conventions hereinafter.

Referring to the example in Figure 2, instead of matching the first "room" (as one might expect), the actual match is much longer. This is due to the use of the greedy quantifier (+). First, the opening quotation mark (") is matched followed by any character (.) at least once (+). As the plus quantifier is greedy, it matches up to the end of the string. After that, the closing quotation mark in the REGEX must be matched. As a consequence, the REGEX engine *backtracks* to a previous state where the last quotation mark in the target string was not matched yet. This process is referred to as *releasing* characters. In the example, the characters already matched are released again one after the other (*i.e.*, first '!', then 's', 'e', 'l', etc.) until a quotation mark is found and an overall match is reported.

As demonstrated in the example above, greedy quantifiers may release characters that were already matched if necessary for an overall match to succeed. However, characters that are *essential* to the quantifier are never released. For example, the plus quantifier (+) in Figure 2 would release

Figure 2: REGEX example: greedy quantifier

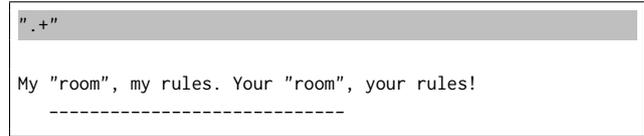
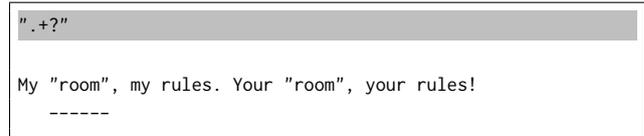


Figure 3: REGEX example: lazy quantifier



all matched characters again if necessary, except the leftmost 'r', because the plus quantifier requires at least one match.

All quantifiers presented in Table 1 can be turned into *lazy* ones by appending a question mark. Their lazy equivalents are therefore `??`, `*?`, `+?`, and `{x,y}?` and they behave in the opposite way to greedy quantifiers, always matching as little as possible. In other words, they only match what is necessary and then hand over control to the next matcher. Only if the next matcher fails, the lazy matcher attempts another match.

Referring to the example in Figure 3, after the first quotation mark (") matches, the lazy dot-matcher matches one character ('r') and then immediately hands over the control to the second '"'-matcher. As the '"' and the next character in the target string 'o' do not match, the control is handed back to the dot-matcher, which matches the 'o'. This procedure repeats until the second quotation mark in the target string is reached and an overall match can be reported.

2.1.2. Non-Backtracking Quantifiers

Similar to the previous section, the greedy quantifiers presented in Table 1 can be turned into *possessive* ones by appending a plus symbol (+). The possessive equivalents are `?+`, `*+`, `++`, and `{x,y}+`. In contrast to matchers with greedy and lazy quantifiers, matchers with possessive quantifiers never backtrack (*i.e.*, never release characters once matched). Figure 4 shows an example where the match result is the same as if a standard greedy quantifier was used. However, for target strings where there is no match (for example, if the target string does not contain a colon), the REGEX engine does not need to backtrack and release already matched characters. That is, a match failure can be reported earlier. This can save memory and increase runtime performance, especially when such a REGEX is applied to many non-matching target strings. This optimization can be applied because, in this example, a colon (:) will never appear among the already matched characters due to the `[a-z]` character class, making backtracking obsolete.

Despite the potential benefits, it is recommended to use possessive quantifiers with caution [14]. For example, in Figure 5, the use of a possessive quantifier leads to a match failure. A greedy quantifier would match all three digits '345' one after the other and then release the '5' again so that the

Figure 4: REGEX example: possessive quantifier

```

^ [a-z]++:
rules: 1) ... 2) ...
-----
    
```

Figure 5: REGEX example: possessive quantifier pitfall

```

\d*+[0-9]
room number 345
(no match)
    
```

character class `[0-9]` can match. However, with a possessive quantifier, this last step does not happen leading to an overall match failure.

2.2. Dynatrace Pattern Language

This section provides details on the most relevant DPL features. It is based on the public DPL documentation [25] and discussions with the original DPL authors. Note that DPL patterns are by default applied exactly once to the target string, *i.e.*, the DPL engine stops execution once it found a match for a given pattern. In addition, matches are only attempted from the beginning of the target string. In other words, every DPL pattern is preceded by an implicit "beginning of line" matcher.

2.2.1. Pattern Syntax

To match literal text in a target string, the required characters must be enclosed in double or single quotes within the DPL pattern. If metacharacters (such as `*`, `[`, and `?`) are to be matched literally, they do not need to be escaped separately, except for the double and single quote and the backslash itself, which must be marked with a preceding backslash (`\`, `'`, and `\\`). Note that literal matchers consisting of multiple letters are considered a single matcher in the DPL. For example, while `abc` represents three consecutive individual matchers in a REGEX, `"abc"` is treated as a single matcher in the DPL.

The DPL provides so-called *export names*, to assign names to parts of the pattern. They are the counterpart to the named capturing groups in REGEXES. To export a matcher, a colon (`:`) is appended, followed by the actual export name. It must start with a letter followed by alphanumeric characters (letters, digits, and underscore). If a period (`.`) appears in the name, the entire export name must be enclosed in quotation marks. Export names can be applied to entire groups, or even individual matchers, and therefore do not necessarily require parentheses. Table 2 shows examples of the most common REGEX features and their equivalents in DPL syntax.

2.2.2. Quantifier Semantics

All the quantifiers listed in Table 1 are also supported by the DPL. Only the question mark (`?`) has slightly different

semantics. It makes the entire matcher with its default quantifier optional, instead of forcing the matcher to match zero times, or exactly once. For example, `DIGIT?` makes the `DIGIT` matcher with its default quantifier `{1,4096}` optional, effectively translating to `DIGIT{0,4096}` (equivalent to `\d{0,4096}` in REGEXES). To exactly match zero or one digit (like `\d?`), the correct DPL pattern is `DIGIT{0,1}`. Similarly, to exactly match one digit (like `\d` in REGEXES), the correct DPL pattern is `DIGIT{1}`. In addition, the DPL supports the range quantifier `{,x}`, which is syntactic sugar for `{0,x}`, *i.e.*, matching minimum 0 and maximum `x` times.

The most critical difference to REGEXES, however, is that despite having the same syntax as greedy quantifiers (\leftrightarrow Section 2.1.1), their behavior is not greedy, but *possessive* (\leftrightarrow Section 2.1.2). The only exception to this is when the quantifiers are applied to the `LD` or `DATA` matchers, which always exhibit *lazy-like* behavior. That is, they match any character up until their succeeding matcher, but (unlike a lazy quantifier in REGEXES), they never expand their initial match.

To emphasize this, the DPL engine *never backtracks* to a previous state, except when evaluating alternatives. Discussions with the original DPL authors revealed that this limitation exists for performance reasons, as backtracking can be an expensive operation. In particular, for certain REGEXES, so-called *catastrophic backtracking* can occur, where the runtime complexity becomes super-linear, also known as *runaway regular expressions* [14, 18, 19, 34, 35].

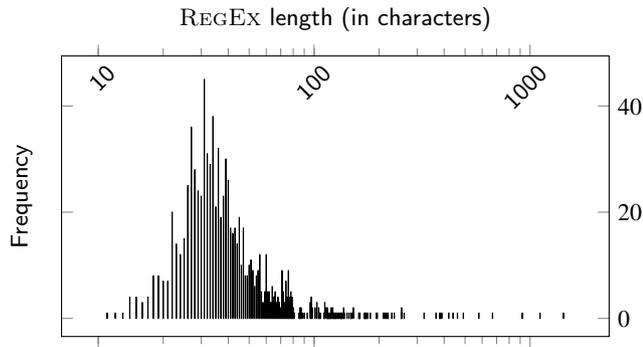
2.2.3. High-Level Matchers

Additionally, the DPL provides a broad range of built-in matchers [36] for frequently occurring information in logs that do not have a direct equivalent in REGEXES. We categorize them into two distinct groups: (i) matchers with a generic scope, *e.g.*, `INT`, `LONG`, `DOUBLE`, and (ii) matchers with a specific scope, *e.g.*, `IPADDR`, `CREDITCARD`, `TIMESTAMP`.

Utilizing these matchers offers a number of advantages. Firstly, they improve the pattern by providing a layer of abstraction, rendering it more readable and maintainable. This not only increases efficiency when writing patterns, but also reduces the risk of errors by eliminating the need to manually create potentially complex patterns. Secondly, they match specific values (*e.g.*, only matching valid credit card numbers), resulting in a more accurate pattern. Thirdly, a data type is associated with the matchers. For example, if a match of type `INT` is exported, the exported value can be processed as an integer (*e.g.*, to easily check whether an HTTP response code is in the range of 200 to 300). In contrast, REGEXES export all matches as strings.

3. Preliminary Study

In this section, we investigate which REGEX features are frequently used in practice, thereby establishing the foundation for our approach. Section 3.1 presents the real-world dataset with 946 individual REGEXES for log parsing. The features used in these patterns are then counted and summarized in Section 3.2.

Figure 6: REGEX length distribution in the real-world dataset


3.1. Real-World Dataset

To identify which REGEX features are frequently used in practice (RQ1), we study a real-world dataset. The dataset was provided by Dynatrace and consists of 5 CSV files with a total size of 3.2 MB. It originates from a large organization in the retail industry which has migrated from Splunk to Dynatrace.⁶ The files contain various *Search Processing Language* (SPL) queries⁷, including REGEXES to parse query results.

For every query, we searched for the `rex` command [38] and extracted the contained REGEX. Because the REGEXES were embedded in both SPL query and CSV format, the REGEXES contained corresponding escapings, which were removed. We obtained 2,701 REGEXES from the original files. After eliminating duplicates, 946 distinct REGEXES finally remained. The REGEXES range from 11 to 1,428 characters in length, with an average length of 56.5 characters. Figure 6 shows the length distribution across all REGEXES. Note that the x-axis has a logarithmic scale.

For every REGEX in the dataset, we counted the features listed in Table 2. To this end, a program written in TypeScript (TS) was implemented using the *regexpp* parser library [39] to analyze the REGEX. First, match mode modifiers were removed so that the REGEX can be parsed by the *regexpp* library. Further, the following optimizations were applied:

- Simplify character classes that contain a single element. Replace `[x]` by `x` if `x` is a single literal matcher or class shorthand.
- Simplify negated character classes that contain a single class shorthand. Replace `[^\d]` by `\D` and vice-versa (`[^\D]` by `\d`) for all three kinds of class shorthands (`\d`, `\w`, and `\s`).

Next, each REGEX was parsed, and the resulting abstract syntax tree (AST) was then traversed recursively. For every intermediate or leaf node, the respective REGEX feature was

⁶Please note, for reasons of confidentiality, neither the original dataset nor the extracted REGEXES can be made publicly available. The examples presented in this work were selected to not contain sensitive information to ensure data privacy compliance.

⁷Search Processing Language (SPL) queries are used to search, filter, and manipulate data within the Splunk platform [37].

counted. For alternatives, we counted the number of pipe characters (e.g., `(a|b)` counts as 1, `(a|bc|d)` counts as 2 and so forth).

3.2. Results

After analyzing all REGEXES, the total feature count was obtained by summing up the counts within each REGEX. Table 2 presents the feature counts over all REGEXES. Column *Total #* shows the total number of occurrences per feature in the entire dataset, while *Affected #* counts the number of REGEXES where the respective feature occurs at least once. The table is sorted by the last column *Affected %*, which shows the percentage of affected REGEXES, calculated by dividing the number of affected REGEXES by the total number of REGEXES (946 entries). For every feature, a REGEX example is given along with its corresponding DPL counterpart. The ten most common REGEX features are separated from the remaining ones by a horizontal line. REGEX features that do not appear in Table 2 (such as match boundaries, backreferences, and comments) did not appear in the entire dataset. One exception are mode modifiers which had to be excluded as mentioned above.

The prevalence of named capturing groups can be explained by their use in SPL for data extraction [9]. The three REGEX features marked as \times in Table 2, namely greedy quantifier, lazy quantifier, and quantified named capturing group, hinder a direct automatic translation. They are not supported by DPL. Due to the low number of occurrences, we ignore quantified named capturing groups. For the other two, we observed that at least one of these quantifiers occurs in 930 REGEXES. Three more REGEX features marked as \diamond , namely character class, negated character class, and non-word boundary, are not directly supported by DPL, but can be emulated.⁸

Answer to RQ1

The top-10 REGEX features include named capturing groups, greedy quantifiers, (escaped) literals, (negated) character classes, non-negated class shorthands, and the dot-matcher.

The results from this study mean that only about 1.7% of all REGEXES from the original dataset can be converted directly into DPL patterns. The following section presents our REPTILE approach that aims to increase this proportion.

4. REPTILE Approach

This section presents our REGEX converter approach REPTILE (Regular Expression Pattern Translation Into Language Equivalents). Figure 7 shows an overview of REPTILE.

⁸Character classes are compatible to DPL, with the exception of class shorthands within character classes. For example, the REGEX `[\d\w]` can be converted to `[0-9a-zA-Z_]` or `(DIGIT{1}|WORD{1})` in DPL. Non-word boundaries (`\B`) can be emulated with lookarounds, which are supported in DPL.

Table 2

Frequency of REGEX features (derived from Friedl [14]) in the real-world dataset; REGEX features unsupported in DPL are marked (✗); REGEX features where conversion is possible with workarounds are marked (❖); remaining REGEX features can be directly converted to DPL.

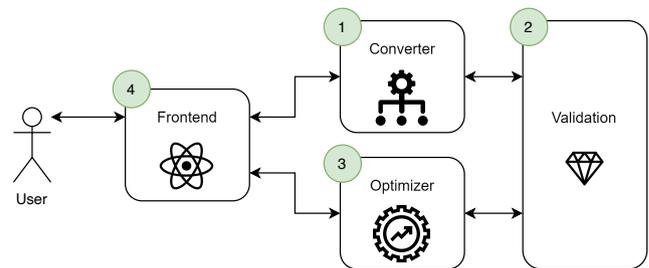
Feature Name	REGEX	DPL	Total #	Affected #	Affected %
Named capturing group	(?<name>abc)	"abc":name	1517	945	99.9%
Greedy quantifier (✗)	a*	✗	2640	893	94.4%
Literal character	abc	"abc"	15983	891	94.2%
Digit matcher	\d	DIGIT{1}	825	384	40.6%
Character representation	\n	LF	1142	370	39.1%
Dot-matcher	.	LD	673	328	34.7%
Character class (❖)	[abc]	[abc]	402	233	24.6%
Space matcher	\s	SPACE{1}	627	185	19.6%
Word matcher	\w	WORD{1}	503	155	16.4%
Negated character class (❖)	[^abc]	[^abc]	282	143	15.1%
Lazy quantifier (✗)	a*?	✗	177	63	6.7%
Capturing group	(ab)c	("ab")"c"	108	53	5.6%
Line start	^	BOS	45	45	4.8%
Alternative	a bc	("a" "bc")	85	41	4.3%
Quantified group	(\s\w)*	ARRAY{SPACE{1} WORD{1}}*	72	40	4.2%
Non-space matcher	\S	NSPACE{1}	63	33	3.5%
Non-capturing group	(?:abc)	("abc")	44	30	3.2%
Non-word matcher	\W	[^a-zA-Z0-9]	13	10	1.1%
Line end	\$	EOS	8	8	0.8%
Optional group	(abc)?	("abc")?	42	7	0.7%
Non-digit matcher	\D	[^0-9]	9	6	0.6%
Positive lookahead	(?=abc)	>>"abc"	3	3	0.3%
Quantified named capturing group (✗)	(?<name>abc)*	✗	4	2	0.2%
Non-word boundary (❖)	\B	❖	1	1	0.1%

In the first phase, the user submits a REGEX through the Frontend (4), which is then forwarded to the Converter component (1), where the basic rule-based conversion from REGEX to DPL happens. Next, the converted DPL pattern is checked for correctness using the Validation (2) component. The converted pattern is presented in the Frontend (4). At this point, the user can already decide to accept the DPL pattern as final result.

In an optional second phase, optimization of the DPL pattern can be initiated. In this case, the basic DPL pattern is forwarded to the Optimizer (3) component, which detects potential high-level DPL matchers. In the Frontend (4), the user must then manually decide for each matcher whether the suggestion should be applied. Optionally, the optimized pattern can be validated again. The following sections provide detailed information on each individual component.

4.1. Rule-Based Conversion (Step 1)

The Converter component utilizes the *regexpp* parser library [39] to parse the input REGEX (similar as in Section 3). The resulting AST is then traversed depth-first. Every conversion step handling one AST node type incrementally extends the DPL pattern result. For most REGEX features, a one-to-one mapping to DPL can be implemented, as shown in Table 2. However, greedy and lazy quantifiers require special consideration. To effectively convert greedy and lazy quantifiers from a REGEX to possessive quantifiers in DPL,

Figure 7: Overview of the REPTILE approach


it is necessary to identify situations where no backtracking is required. In the following, we present these situations and our conversion strategies.

4.1.1. Greedy Quantifiers

We identify three cases where greedy quantifiers do not require backtracking and therefore can be converted to DPL.

Fixed Greedy Quantifier (FGQ). The first case of quantifiers without backtracking are quantifiers with fixed repetition, such as $\{x, x\}$ or $\{x\}$ in short. As mentioned in Section 2.1.1, greedy quantifiers may only release non-essential characters. For example, in the case of the quantifier $\{2, 4\}$, only the first two matched characters are essential, and the following two may be released again if needed. Similarly, for the plus quantifier $\{+\}$, the first matched character is

Figure 8: Safe greedy quantifier: last matcher in a REGEX

```
method=[A-Z]*
method=POST, endpoint=https://...
-----
```

Figure 9: Safe greedy quantifier: followed by non-intersecting literal matcher

```
\d{1,3}x
789
(no match)
```

Figure 10: Unsafe greedy quantifier: followed by intersecting character class

```
\w+[a-z]
Hello-Muehlviertel!
-----
```

Figure 11: Unsafe greedy quantifier: followed by optional matcher and character class

```
\w+\s?[a-z]
Hello-Muehlviertel!
-----
```

essential, all following ones may be released again. In the quantifier with fixed repetition $\{x\}$ however, *all x characters are essential*, and therefore, there is no backtracking, *i.e.*, no characters are released.

Last Greedy Quantifier (LGQ). The second case, where no backtracking can happen, is when the greedily quantified matcher appears last in the REGEX (or last in any top-level alternative). For example, in Figure 8, once the 'method=' part matched, the following $[A-Z]^*$ part greedily matches uppercase letters until either (i) the end of the target string is reached, or (ii) a character is reached which does not match the expression, *i.e.*, is not an uppercase letter. This example shows the latter case. However, in either case, the $[A-Z]^*$ never has to release any already matched characters again because no matcher comes afterwards that requires one of the matched characters.

Non-Intersecting Greedy Quantifier (NGQ). The third case where a matcher with a greedy quantifier can be safely converted is if its succeeding matcher matches different characters. That is, the languages defined by a matcher and its successor do not intersect. In set mathematics, detecting this condition is called the *intersection (non-)emptiness problem*.

In the implementation of REPTILE, we follow an existing approach [40] and use the *greenery* library [41] to check whether two matchers intersect. First, consider the simplest case: a quantified matcher succeeded by a literal matcher as shown in Figure 9. Here, the x attempts to match, after the greedy $d\{1,3\}$ reaches the end of the target string. As the x requires exactly one match, the digit matcher releases the last matched character 9. The x attempts to match the 9 and fails, resulting in an overall failure. It is important to realize that this backtracking step is always unnecessary, no matter to which target string the example REGEX is applied to. The reason is that the two languages defined by $d\{1,3\}$ and x do not intersect. While $d\{1,3\}$ matches any numbers between 0 and 999 (including numbers with leading zeros, such as 001), the literal matcher x solely matches the lowercase character x . Hence, no matter which digit is released by the digit matcher, it can never be consumed by the literal matcher x . In other words, the result does not change, when the greedy quantifier is replaced by a possessive one, making it safe to convert to DPL. This insight is summarized in the following observation:

Observation. *Given a REGEX P , let $P = uv$, where u and v are subpatterns of P . Let u be greedily quantified, and v be non-optional. Let $L(u)$ and $L(v)$ be the languages defined by u and v , respectively. If $L(u) \cap L(v) = \emptyset$, then the quantifier of u can be replaced by a possessive quantifier without affecting $L(P)$.*

Figure 10 shows an example where conversion is unsafe. After the \wedge^+ has matched `Hello`, it passes the control to the final $[a-z]$. As this last matcher requires exactly one character, it forces the initial \wedge^+ to release one character, finally matching the `o`. In summary, as the greedy \wedge^+ and the succeeding matcher $[a-z]$ intersect, backtracking may be necessary. Hence, the REGEX cannot be safely converted to DPL.

The following special cases apply. In case the successor is a group, the first element must be chosen from the group and checked for intersection. If the group contains alternatives, intersection must be checked for the first element of every alternative. In other words, the first element of every alternative must not intersect with the current matcher. Note that *successor* and *first element* refer to the next non-optional matcher after the current matcher. The reason is that optional matchers never cause the predecessor matcher to release characters that have already been matched, as they do not require any characters themselves. Consider the example in Figure 11. Although \wedge^+ and $\s?$ do not intersect, the result does not differ from the unsafe example in Figure 10. Because $\s?$ is optional, and the target string does not contain a whitespace character, this matcher is skipped. The last matcher in turn initiates backtracking, thus forcing the initial \wedge^+ to release one character. In summary, because the greedy \wedge^+ and the next non-optional matcher $[a-z]$ intersect, backtracking may be necessary and the REGEX cannot be safely converted to DPL.

Figure 12: Safe lazy quantifier: followed by non-intersecting literal matcher

```
\d+?x$
78xx
```

Figure 13: Unsafe lazy quantifier: followed by intersecting character class

```
\w+?[a-z]
Hello-Lavanttal!
--
```

Figure 14: Safe lazy quantifier: dot-matcher followed by exactly one matcher

```
!.+?!
Hello! Zillertal!
-----
```

4.1.2. Lazy Quantifiers

Most strategies for converging greedy quantifiers presented in the previous section also apply to lazy quantifiers.

Fixed Lazy Quantifier (FLQ). As explained in Section 2.1.1, lazy quantifiers expand their match if necessary for an overall match to succeed. However, if a quantifier has fixed repetition ($\{x,x\}$ or $\{x\}$) the quantifier can never expand its match beyond the initial match of length x . Hence, a matcher with such a quantifier can be safely converted to its possessive DPL counterpart.

Non-Intersecting Lazy Quantifier (NLQ). Next, a matcher with a lazy quantifier can be safely converted if it does not intersect with its next non-optional succeeding matcher as shown in Figure 12. Here, $\backslash d+?$ first matches the minimum number of required characters, namely the character 7. Control is passed on to the x matcher, which cannot match the next character in the target string (8), forcing $\backslash d+?$ to expand its match by consuming 8. Then, x consumes the first x in the target string. After that, the end-of-line matcher (\$) fails, forcing the x to release the character again. Now, $\backslash d+?$ is asked to expand its initial match, which fails, resulting in an overall fail. It is important to realize that the lazy quantifier always expands its match until the succeeding x can match the first time, but never beyond, no matter to which target string the example REGEX is applied to. The reason is that the two languages defined by $\backslash d+?$ and x do not intersect. Hence, if the literal matcher releases its matched character, it can never be consumed by the digit matcher. In other words, the result does not change, when the lazy

quantifier is replaced by a possessive one, making it safe to convert to DPL. In conclusion, the observation from the previous section is not limited to greedy quantifiers, but also applicable to lazy quantifiers.

However, when the matchers intersect, the conversion cannot be safely performed, as shown in Figure 13. After $\backslash w+?$ consumes the first character H, $[a-z]$ consumes the second one, resulting in an overall match. The possessive DPL counterpart would consume all characters of Hello until there is no character for the $[a-z]$ left to match, resulting in an overall fail.

Last Lazy Quantifier (LLQ). If a lazy quantifier appears before an end-of-line matcher (\$), it always matches until the end of the target string. Therefore, this conversion is safe as well. However, if the matcher appears at the very end of the pattern, different rules apply. Because no matcher comes afterwards, the last matcher never has to expand beyond its initial match. In other words, it always matches the minimum required number of characters if possible. That is, lazy quantifiers $\{x,y\}?$ at the end of a pattern must be converted to $\{x\}$. Consequently, $+?$ translates to $\{1\}$, and $*?$ translates to $\{0\}$. The latter case implies that the entire matcher never consumes any characters, so it can be omitted.

Last Successor Lazy Quantifier (SLQ). Lastly, one special case applies to the dot-matcher in combination with a lazy quantifier, such as $.+?$ and $.*?$. If such a matcher is followed by exactly one matcher, *i.e.*, its successor is the last matcher in the pattern, it can be safely converted to the *lazy-like* counterpart LD . As shown in Figure 14, the REGEX behaves the same way as its DPL equivalent $LD+ "!"$ would. Note that any additional matcher may break this behavior. For example, adding a EOL (respective \$) matcher at the end, would require the LD to expand its match, which is not supported, resulting in an overall fail. Note that multiple literal matchers are considered one single matcher in the DPL, as mentioned in Section 2.2.1. Hence, this strategy is also applicable to patterns where a dot-matcher is followed by multiple literal matchers, such as $.+?abc$, which translates to $LD+ "abc"$.

4.2. Pattern Validation (Step 2)

The original REGEX and the generated DPL pattern are considered semantically equivalent only if their defined languages are equivalent. That is, the DPL pattern must accept all strings that are accepted by the REGEX and reject all strings that are rejected by the REGEX. However, when dealing with REGEXES, it is often impossible to enumerate all strings of the defined language, if it is infinite (*e.g.*, when unbound quantifiers such as $*$ or $+$ occur). One way to test the equivalence of two regular expressions is to compare their minimal deterministic finite automata (DFA) representations [42]. However, this cannot be used in our approach, because no algorithm exists that converts DPL to DFA.

To overcome this, we adopt the approach proposed in [43], randomly generating a fixed number of strings based

on the original REGEX. These generated strings then serve as positive test cases for the DPL pattern. Negative test cases can be obtained by generating matching strings based on the REGEX's complement. Those negative tests pass if they are *not matched* by the converted DPL pattern. Although such test cases cannot formally proof semantic equivalence, they provide strong empirical evidence (similar to unit tests). In the remainder of the paper, we name a pattern translation from REGEX to DPL that passed these criteria as a "safe" translation.

All test cases are finally matched against the converted DPL pattern. A positive test case is considered to be passed, only if both (i) it matches the string entirely (ii) it extracts the same field values as the original REGEX.

4.3. Pattern Optimization (Step 3)

After a REGEX was converted to DPL and validated, optimization of the DPL pattern can be initiated. That is, potential high-level DPL matchers (\leftrightarrow Section 2.2.3) are detected.

4.3.1. Optimization Objective

One potential solution to detect DPL's high-level matchers is to map each of them to a corresponding REGEX and search for this sub-pattern in the original REGEX. However, it is unlikely that an exact match can be found.

Figure 15 presents two illustrative REGEX examples. Both examples extract an IP port, which can be matched by INT in DPL. Line 2 illustrates a variant that precisely matches one to five digits, which is less than what would be matched by the INT matcher. Line 5 depicts a variant that matches an arbitrary number of digits, which is more than what would be matched by the INT matcher. Nevertheless, it is desired to use the INT matcher in both cases for the reasons mentioned in Section 2.2.

Figure 15: REGEX matching an IP port

```

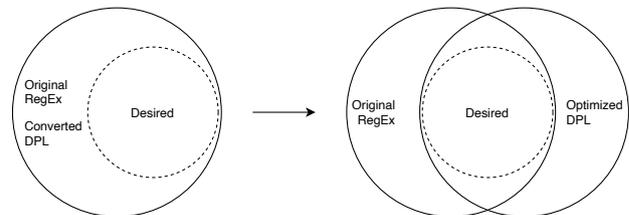
1 // matching less than INT
2 port\s(<?ip_port>\d{1,5})
3
4 // matching more than INT
5 port\s(<?ip_port>\d*)
    
```

Purely algorithmic detection based on the defined language of the high-level matcher therefore does not seem feasible. Instead, other factors shall be included in the detection:

- The name of a named capturing group. For example, an export name `src_addr` may indicate that the group extracts an IP address. Note that these names are user-defined and may appear in languages other than English.
- The context of the REGEX. For example, if a REGEX contains the literal matcher `HTTP/1\.`, an export name `reponse_code` can most likely be interpreted as an HTTP response code.

We hypothesize that a machine learning (ML) based approach could yield useful results to effectively identify the context of a pattern and the dependencies between the matchers to suggest potential high-level DPL matchers. For each prediction, the user must manually review the suggestions and confirm their suitability for the intended use. This is necessary because the semantics of the DPL pattern most likely changes when using high-level matchers. To emphasize this, Figure 16 compares the language of an original REGEX with the language of its converted DPL pattern before and after detection of high-level matchers. Even in the event that the languages of both REGEX and DPL pattern are semantically equivalent (see left part of Figure 16, *i.e.*, a "perfect" translation), high-level matchers can still be employed. After this, the defined languages of the REGEX and DPL pattern may differ (right part of Figure 16). However, this is acceptable as long as the optimized DPL pattern continues to include the *desired* language, which represents the REGEX author's true intent. Note that a REGEX often is an approximation of this *desired* language [20].

Figure 16: Language comparison of a REGEX and DPL pattern before (left) and after (right) the optimization step



4.3.2. Matcher Prediction

To detect the high-level DPL matchers as explained above, the basic DPL pattern as a result of the rule-based translation is sent together with the problem and task to GPT-4.

One advantage of GPT-4 is that the model is already pre-trained [44] and can understand inputs in natural language as well as programming languages and deliver results without fine-tuning. OpenAI does not provide specific information about the training data used. Instead, it is vaguely stated that it includes "publicly available data (such as internet data)" [44]. Consequently, the training dataset may also contain the public DPL documentation [25]. However, our initial interactions with the model have shown that GPT-4 is not able to generate valid DPL patterns from a target description alone (*e.g.*, "Generate a DPL pattern that matches..."). Therefore, we chose to use the *zero-shot prompting* strategy describing the optimization task in detail. This technique can lead to more robust results compared to prompts containing demonstrations [45].

The prompt consists of the following four sections: First, the basic DPL pattern is presented as a whole and separated into *fragments*. We use the term *fragments* to refer to the individual matchers, alternatives, and groups on the highest level of a pattern. For example, `"a"*"b"("c"|"d")?` consists

Figure 17: REPTILE’s frontend showing a converted pattern including high-level matcher suggestions after pattern optimization



of three fragments "a"*, "b", and ("c"|"d")?. Then, the optimization task is described including a list of high-level matchers which can be suggested for each fragment. To reduce complexity (also in regards to evaluation), the list of included high-level matchers is limited to the selection of five specific and generic matchers listed below⁹. These matchers were selected, because we observed that real-world application logs often contain recurring information such as IP addresses, timestamps, IP ports, HTTP response codes, etc. which are covered by these DPL matchers.

- IPADDR (matching IPv4 and IPv6 addresses)
- INT, LONG, DOUBLE
- TIMESTAMP (with default format yyyy-MM-dd HH:mm:ss)

Next, criteria are defined, based on which matchers can be predicted, namely a matcher’s export name and its accepted language. It is explained that parentheses can be omitted if an entire group is replaced by a matcher and where the export name must be attached. Lastly, the response format is restricted to JSON and the expected response schema is defined. Each prompt is preceded with the following message: "You act as a backend suggesting optimizations for the DPL (Dynatrace Pattern Language) responding in plain JSON." The full prompt is provided in a public GitHub repository [46].

4.4. REPTILE Frontend (Step 4)

The Frontend component, implemented solely for validation purposes, displays conversion and validation results. It also facilitates the review process during pattern optimization. Figure 17 shows a screenshot of REPTILE’s frontend¹⁰. The REGEX for conversion can be entered in the topmost input field. Then, the "Convert" button in the top right initiates the rule-based conversion. After conversion is finished,

⁹A complete list of built-in matchers is available online [36].

¹⁰We also provide a screencast to illustrate the tool and its frontend <https://video.dynatrace.com/watch/oxUekmeEehmGHTAZ29d8HN?>

the generated DPL pattern is shown in the input field on the bottom where the result can be modified if necessary. Above, the results of the verification and validation steps are shown as explained in Section 4.2. Test results can be analyzed in detail, by clicking the "... tests passed" chip, which opens a sheet overlay containing the test results.

For visibility, the pattern fragments (↔ Section 4.3.2) are shown in the central component labeled "Pattern Conversion". When the mouse cursor is moved over a fragment, both the fragment and its REGEX or DPL counterpart are highlighted. This way, the user can trace how each fragment of the REGEX was converted into which fragment of the DPL pattern. This is most relevant in case of a best-effort conversion (shown as an "unsafe" conversion in the UI). That is, the affected greedy or lazy quantifier is transformed into its corresponding possessive DPL quantifier. In this case, the respective fragment’s background color is changed to yellow. Moreover, the reason why a conversion could not be safely done is shown for the respective fragment, e.g., because the quantified matcher intersects with its successor. Based on this information, the user can decide whether to accept or modify the pattern.

Optionally, the user can initiate the optimization of the DPL pattern. Each suggested alternative is then rendered above its respective fragment. The user can decide which alternative to choose by clicking on the fragments. The final DPL pattern in the bottom line is updated after every selection. After each update, the syntax is checked automatically. Generating and executing the test cases can be triggered manually via "Run tests" or "Rerun" buttons respectively.

4.5. Implementation Details

We created REPTILE as a custom React single-page application (SPA) that uses two services from the Dynatrace environment to validate DPL patterns. To verify the DPL patterns’ syntax, the *Query Assistance Service* is used. To validate the semantics, the pattern is matched against the test

cases described in Section 4.2 with the help of the *Pattern Developer Service*. Furthermore, the *App Toolkit* [47] facilitates rapid app development, setup, and deployment, while the *Strato Design System* [48] offers pre-built UI components for creating a contemporary and consistent user interface. We currently cannot provide a publicly available version of the prototype because it is bound to internal services of Dynatrace. However, in the future we will work on minimizing the dependency on the Dynatrace environment to facilitate the usage of our prototype.

To detect the high-level DPL matchers as explained in Section 4.3, a dedicated GPT-4 instance without fine-tuning is used. It is accessed via the *Azure OpenAI Service* [49]. To detect if the defined languages of an element and its successor intersect (\leftrightarrow Section 4.1), we use the *greenery* library [41]. Because this library is written in Python, it cannot be directly used within REPTILE. To overcome this limitation, we deployed *greenery* as an AWS Lambda serverless function. This library is also used to obtain a REGEX's complement necessary to generate negative test cases (\leftrightarrow Section 4.2). The *reregexp* library [50] is used to generate random strings based on the original REGEX and its complement, which serve as positive and negative test cases as described in Section 4.2.

5. Evaluation

This section evaluates the REPTILE approach. First, Section 5.1 assesses the proportion of safely convertible REGEXES. Please note, as described above, we use the term "safe" conversion if a REGEX can be converted into DPL while retaining the original semantics. Second, Section 5.2 evaluates the accuracy of REPTILE's pattern optimization.

5.1. Rule-Based Conversion

To answer RQ2, this section assesses the efficacy of our approach. To this end, we utilized the real-world dataset from Section 3 to assess the number of REGEXES affected by our conversion strategies. The program for counting the REGEX features was extended by a new feature category for every conversion strategy presented in Section 4.1.

Tables 3 and 4 show the coverage of the quantifier conversion strategies presented in Section 4.1. Overall, 2,640 greedy quantifiers were found in 893 out of 946 REGEXES of our industrial dataset. The majority of greedy quantifiers, namely 1,546 (58.56%), are NGQs occurring in 386 REGEXES. Note, while LGQs did not occur most frequently, they affected more REGEXES, namely 550. Regarding greedy quantifiers, 658 (*i.e.*, 893 - 235) out of 893 (= 73.7%) affected REGEXES were safely converted with REPTILE. 235 REGEXES contained at least one greedy quantifier for which REPTILE does not provide a conversion strategy.

In contrast to greedy quantifiers, only 177 lazy quantifiers were found in 63 out of 946 REGEXES. The majority of them, namely 41, are SLQs occurring in 41 REGEXES. Our

Table 3
Greedy quantifier conversion results

Type	Total #	Affected #
Fixed Greedy Quantifier (FGQ)	167	94
Last Greedy Quantifier (LGQ)	558	550
Non-Intersecting Greedy Quantifier (NGQ)	1546	386
Remaining	369	235
Total Greedy Quantifiers	2640	893

Table 4
Lazy quantifier conversion results

Type	Total #	Affected #
Fixed Lazy Quantifier (FLQ)	0	0
Non-Intersecting Lazy Quantifier (NLQ)	8	6
Last Lazy Quantifier (LLQ)	0	0
Last Successor Lazy Quantifier (SLQ)	41	41
Remaining	128	21
Total Lazy Quantifiers	177	63

industrial dataset did not contain any FLQs or LLQs. Regarding lazy quantifiers, 42 (*i.e.*, 63 - 21) out of 63 (= 66.7%) affected REGEXES were safely converted with REPTILE.

Table 5 shows the results for all conversion strategies applied in combination. Column *Affected #* corresponds to the number of REGEXES covered. *Affected %* reports the same proportion relative to the dataset size (946 REGEXES). The results show that REPTILE's rule-based conversion increased the number of safely converted REGEXES by 681 to 697 (= 16 + 681) or 73.7% safely converted REGEXES.

Due to an occurrence of a quantified named capturing group, conversion was syntactically impossible for only 2 REGEXES. For the remaining 26.1% of the REGEXES, REPTILE provided a best-effort conversion. The majority of these (17.2%) cannot be safely converted, because they contain a quantified dot-matcher (either greedy or lazy). The remaining 8.9% contain some other quantified matcher, which in all cases are observed to be greedy. Note that these REGEXES may also contain dot-matchers, but additionally contain at least one other quantified matcher.

The correctness was validated for the 697 (= 16 + 681) converted REGEXES, using randomly generated samples as described in Section 4.2. Due to a dependency on the *greenery* library (as discussed in Section 4.5), negative test case generation was only conducted for a subset of REGEXES. The reason for this is that *greenery* is rigorous in its handling of escapings. While in PCRE, for example, both $\%$ and $\%$ are recognized as literal matchers, *greenery* only allows the latter variant. However, not all special cases are documented in the library. To work around this issue and also save resources, we created negative tests for 200 randomly selected REGEXES which are accepted by *greenery*.

We performed two test runs with 500 test cases each, resulting in 1,000 samples per REGEX. All 897,000 generated test cases (697 * 1,000 = 697,000 positive and 200 * 1,000 = 200,000 negative test cases) were reported to be successful. That is, for positive tests, both the overall matches and their

Table 5
Conversion results for all conversion strategies applied in combination

Type	Affected #	Affected %	Combined %	Conversion
Initially converted	16	1.7%	73.7%	safe
Additionally converted	681	72.0%		
Dot-matcher with greedy or lazy quantifier	163	17.2%	26.1%	best-effort
Remaining greedy quantifiers	84	8.9%		
Quantified named capturing groups	2	0.2%	0.2%	not possible
Total	946	100.0%	100.0%	-

individual captured values were found to be equivalent. For the negative tests, no match was reported as desired. In conclusion, these results indicate full semantic congruence and, therefore, correct conversion of *all* converted REGEXES.

Answer to RQ2

In a real-world dataset containing 946 REGEXES, REPTILE increased the number of safely converted REGEXES (that is, automatically without human judgment) from initially 1.7% to 73.7%.

5.2. Pattern Optimization

This section presents the evaluation of the correctness of REPTILE’s DPL pattern optimization to answer the final research question RQ3.

5.2.1. Evaluation Datasets

For this evaluation, we require a set of REGEXES and actual log entries that represent the ground truth to check whether the optimized DPL patterns correctly match the corresponding parts of a log entry. For that, we could not use the REGEXES from the previous evaluation because, due to confidentiality reasons, we did not have access to the actual log files from our industrial partner.

To address this issue, we first selected a set of technologies and then searched the internet for corresponding logs and REGEXES to parse them. We limited the technologies to logs, such as from Apache, NGINX, AWS, for which *DPL Architect* [51] provides built-in DPL patterns. For each technology, we entered "[technology] logs" on Google and perplexity¹¹ to search the internet for log files from the respective technology. Next, we entered "[technology] regex" to search the internet for REGEXES with which respective log files can be parsed. In total, we collected 23 REGEXES from 13 technologies, which are enumerated in Table 6. The column *Logs #* shows the overall number of log entries found for each technology. The column *Sources* provides information on the logs’ origins. The majority of REGEXES originates from the *Regex101 community* [52] and the official *ChaosSearch* documentation [53]. Their length ranges from 121 to 787 characters, with an average length of 271.6 characters. We provide the REGEXES and links to the corresponding log files in a public GitHub repository [46].

¹¹<https://www.perplexity.ai/>

For each technology, we first combined all logs into a single file. Next, we executed each REGEX on the corresponding log file to collect all log entries that match the REGEX. These log entries represent the positive test cases. Conversely, negative test cases were obtained by filtering the corresponding log file for non-matching log entries. The columns ϵ^+ and ϵ^- in Table 6 show the number of positive and negative test cases respectively. Note, for REGEXES with more than 1,000 test cases, the log entries were sampled randomly and limited to 1,000 (marked as * in Table 6).

Furthermore note, for seven REGEXES (P1a–P1d, P7a, P8a–P8b), we observed that some of the corresponding log entries contain optional elements. For example, some log entries contain dashes in places where IP addresses are expected. These are matched by the original REGEX and its corresponding DPL pattern (*e.g.*, via LD or NSPACE), but not by the high-level matcher IPADDR, because a dash is not a valid IP address. Therefore, we removed these log entries and, consequently, test cases from the datasets leading to the number of positive test cases actually used for the evaluation presented in column ϵ_S^+ in Table 6.

5.2.2. Experimental Set-Up

Using the dataset from above, we first applied REPTILE’s rule-based conversion to convert each of the 23 REGEXES to a DPL pattern and optimize it with REPTILE’s DPL optimization (\hookrightarrow Section 4.3). As already explained in that section, we used an OpenAI’s GPT-4 instance via Azure OpenAI Service and our zero-shot prompting strategy for the optimization.

Next, for every optimized DPL pattern we obtained the fragments (\hookrightarrow Section 4.3.2) for which a prediction of a high-level matcher was made. For each predicted high-level matcher in a DPL pattern, we classified it as True Positive (TP) or False Positive (FP) by replacing the fragment by the predicted matcher and execute the pattern on the test cases listed in Table 6. If *all* test cases passed, the prediction was categorized as TP, otherwise as FP. An exception to this is the *TIMESTAMP* matcher. Because this matcher is used with its default date format (\hookrightarrow Section 2.2.3), a prediction was categorized as TP if it was predicted for any fragment matching a timestamp (= *fragment hit*), even if the default format was incorrect.

The remaining fragments of a pattern were manually categorized as True Negatives (TN) or False Negatives (FN) as follows. A fragment for which the matcher can be applied

Table 6

23 REGEXES with their corresponding positive (ϵ^+), sanitized positive (ϵ_S^+), and negative (ϵ^-) test cases; entries marked as * are randomly sampled and limited to 1,000

REGEX	Technology	Sources	Logs #	ϵ^+	ϵ_S^+	ϵ^-
P1a	Apache Access	[54, 55]	95,332	*1,000	981	-
P1b				*1,000	973	470
P1c				*1,000	982	*1,000
P1d				*1,000	986	*1,000
P2a	Apache Error	[54]	67,456	*1,000	1,000	*1,000
P2b				110	110	*1,000
P3a	NGINX Access	[52]	51,462	*1,000	1,000	51
P3b				*1,000	1,000	51
P3c				*1,000	1,000	-
P3d				*1,000	1,000	-
P4a	OpenSSH	[52, 56]	2,035	525	525	*1,000
P5a	AWS Cloudfront	[57]	76	10	10	66
P5b				76	76	-
P5c				76	76	-
P6a	AWS Route 53	[58]	5	5	5	-
P7a	AWS S3	[59]	10	10	8	-
P8a	AWS VPC Flow	[60–62]	24	24	21	-
P8b				24	19	-
P9a	Core DNS	[52, 63, 64]	15	15	15	-
P10a	Log4j	[52]	42	42	42	-
P11a	NGINX Error Log	[52]	27	27	27	-
P12a	SV.NET Service Bus	[52]	22	22	22	-
P13a	HDFS Audit Log	[52]	31	31	31	-

but was not predicted was categorized as FN. All other remaining fragments were categorized as TN. These are the fragments for which the matcher could not be applied and was also not predicted by REPTILE. The manual categorization was performed by the first author and the results were validated by a senior software engineer at Dynatrace, who is responsible for the development and maintenance of the DPL. Conflicting cases were discussed by both of them and agreement was reached for all cases.

Consider the DPL pattern `LD*:ip ":" WORD+:msg` that consists of the three fragments `LD*:ip`, `":"`, and `WORD+:msg`. For the sake of argument, assume that `IPADDR` was predicted for the first and last fragment during optimization. This results in $TP=1$ (correct prediction of first fragment, assuming all test cases passed) and $FP=1$ (incorrect prediction of last fragment, assuming at least one test case failed). Further, $FN=0$, because there is no additional fragment where `IPADDR` could have been correctly predicted. Lastly, $TN=1$, because the matcher was not predicted for the remaining middle fragment (which is correct).

Based on the TPs, FPs, FNs, and TNs, the precision, recall, F1-score, and Matthews correlation coefficient (MCC) were calculated for each high-level matcher using the formulas below.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

used to obtain the *F1-score* as follows:

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5.2.3. Results

Table 7 shows the results for each high-level matcher across all 23 patterns as described before. The column Σ shows the total number of fragments in our dataset, namely 579. Referring to the results, we observe high precision values of more than 0.85 across all matchers. This is due to the low number of false positives (FPs) ranging between 0 and 5 for the five high-level matchers. This means that when a prediction is made by REPTILE, it is correct in most cases.

The recall values indicate that for most fragments for which a high-level matcher is possible, such a matcher is also predicted by REPTILE. We emphasize that a recall of 1.00 was achieved for the three matchers `IPADDR`, `LONG`, and `TIMESTAMP`, indicating that all optimization potential was utilized. A relatively low recall of 0.71 is only observed for the `DOUBLE` matcher. However, this can be attributed to the limited number of 7 fragments for this matcher, potentially leading to less representative results.

Looking at the F1-score and MCC values in Table 7, we observe values of 0.91 and higher for 4 high-level matchers, whereas the best result (0.98 F1-score and MCC) was obtained for `TIMESTAMP`. Also the F1-score and MCC of 0.83 and 0.84, respectively, for `DOUBLE` indicate good prediction performance of REPTILE's pattern optimization.

Answer to RQ3

Optimizing 23 real-world patterns, REPTILE shows an average F1-score of 0.91 and an average MCC of 0.92 for high-level matcher prediction across five selected matchers.

6. Discussion

In this section, we first discuss potential implications of our results on practitioners and researchers. Thereafter, we discuss threats to validity of the results from our experiments.

6.1. Implications

The results of our evaluation show that REPTILE can automatically and correctly convert 73.7% of the REGEXES to DPL patterns. Furthermore, for the other 26.5% of REGEXES, it provides a best-effort conversion that need to be reviewed by a user. However, for these best-effort conversions, most manual work is eliminated by REPTILE because the review is only needed for parts/fragments of the REGEXES that could not be converted by one of REPTILE's conversion strategy.

Table 7

REPTILE's pattern optimization results for 23 REGEXES consisting of 579 pattern fragments

Matcher	Σ	TP	FP	FN	TN	Precision	Recall	F1-score	MCC
IPADDR	579	20	3	0	556	0.87	1.00	0.93	0.93
INT	579	43	5	3	528	0.90	0.93	0.91	0.91
LONG	579	11	2	0	566	0.85	1.00	0.92	0.92
DOUBLE	579	5	0	2	572	1.00	0.71	0.83	0.84
TIMESTAMP (fragment hit)	579	23	1	0	555	0.96	1.00	0.98	0.98
Average	-	-	-	-	-	0.91	0.93	0.91	0.92

This clearly shows that REPTILE will save users time and effort.

Furthermore, REPTILE's pattern optimization offers potential improvements that only need to be confirmed or rejected by the user. The results from the evaluation show high precision and recall values across five selected high-level matchers. Consequently, if a fragment in a DPL pattern can be replaced by an adequate high-level matcher, REPTILE predicts it, and most likely it is the correct one. This helps users to create easier-to-understand and maintain DPL patterns with little extra effort.

REPTILE is currently deployed in a Dynatrace internal environment to test its usefulness. Several internal users started using it — within 4 months, we recorded 29 user sessions with an average duration of more than 50 minutes. The tool was also presented at the *Dynatrace Principal Solution Engineer Summit 2024* in Denver. We received positive feedback from *Solution Engineers* that REPTILE is conceived to be highly useful and is expected to greatly support the conversion process. Both pieces of evidence support the benefits of our approach.

Researchers also benefit from the presented results. First, we provide a detailed description of our approach that allows other researchers to adopt it for converting REGEXES to other pattern languages, such as Grok. As mentioned in Section 4.5, we will also work on minimizing the dependency on the Dynatrace environment to allow other researchers to use and extend REPTILE. Finally, we provide a publicly available dataset that currently contains 23 real-world REGEXES and corresponding log files. It can be used to evaluate future extensions of REPTILE or other approaches for converting REGEXES to DPL or other pattern languages.

6.2. Threats to Validity

We identify the following threats to internal and external validity of our findings. Regarding internal validity, we found typos in few REGEXES of the industrial dataset. Figure 18 shows an example where the escaping of the dot in `HTTP\1.1` is missing. This error remained undetected, as the dot-matcher matches *everything*, including the desired literal dot. This resulted in an incorrect count of the matcher as dot-matcher in Section 3, despite its intended purpose of being a literal matcher. However, a manual inspection of a random sample of REGEXES showed that this issue only affects a

few patterns, posing little risk to the validity of the presented results.

Figure 18: REGEX containing a dot with missing escaping

```
HTTP\1.1 (?<RespCode>[0-9]+) (?<RespSize>[0-9]+)
```

Another threat to internal validity concerns the evaluation of the correct conversion of REGEXES to DPL patterns in RQ2. Due to a limitation of *greenery*, negative test cases were not generated for all real-world REGEXES (\leftrightarrow Section 5.1). But, we are confident that the results are still valid because positive test cases were generated and passed for all patterns which suggests that the converted patterns fully cover the language of the original REGEXES. Furthermore, we performed this evaluation with negative test cases for a random sample of 200 REGEXES which are accepted by *greenery*. We view this sample as sufficiently large.

Regarding external validity, we performed the experiments for RQ1 and RQ2 with one industrial dataset from a single organization in a specific industry. Therefore, the results might not generalize to other companies. We mitigated this threat by having our dataset assessed by two experienced solution engineers at Dynatrace. Both found that the dataset contains REGEXES that encompass common use cases for extracting key performance indicators (KPIs) and other business information, similar to those seen from customers in other business domains, such as banking. Furthermore, the evaluation of REPTILE's pattern optimization demonstrated that our approach indeed generalizes to unseen patterns from various other sources. Regarding RQ3, we mitigated this threat by considering REGEXES and log files from 13 different popular technologies.

REPTILE currently supports only the conversion from REGEXES to DPL, which limits the generalizability of the results. However, as mentioned in Section 1, DPL is the log parsing pattern language offered by Dynatrace. And Dynatrace is a leading company in providing software observability solutions [27] used by many companies from different business domains. Therefore, we can safely assume that DPL is a representative pattern language. In addition, REPTILE follows a modular design in which individual components, such as the Converter or the Optimizer (\leftrightarrow Figure 7), can be extended or replaced. This facilitates the adaptation of

REPTILE’s conversion strategies and optimization approach. We plan such extensions as part of our future work, for instance to convert REGEXES to Grok.

Another threat to external validity concerns the evaluation of REPTILE’s pattern optimization in RQ3 (\leftrightarrow Section 5.2). The evaluation was limited to five common high-level matchers while DPL provides more such matchers. We mitigated this threat by selecting the five matchers based on our observations from the real-world dataset. They showed that IP addresses, timestamps, IP ports, HTTP response codes, and other numbers occur frequently in log entries. In future work, we plan to extend our approach to consider further high-level matchers, such as HEXINT or CREDITCARD.

Lastly, REPTILE’s pattern optimization was only performed and evaluated with GPT-4 using one prompting strategy which might limit the generalizability of the results. First, we would like to emphasize that the goal of this study was not to find the best-performing LLM and prompting strategy, but to assess the feasibility of LLM-based pattern optimization. Second, we chose GPT-4 because it showed good performance in solving many other software engineering tasks [65]. Future work will be concerned with exploring other LLMs and deep learning algorithms in addition to model fine-tuning [66] and other prompt engineering strategies, such as *few-shot prompting* [45] or *chain-of-thought prompting* [67].

7. Related Work

Due to the breadth of its use, the body of research on regular expressions is extensive and in its entirety beyond the scope of this work. Thus, we focus here on approaches with similar intent — the generation or conversion from and to regular expressions within a limited set of applications.

The conversion of regular expressions has been of interest for a longer time starting with Glushkov’s automata [68] to convert a REGEX into nondeterministic finite automata (NFA) [69–71]. In addition, the conversion to parsing expression grammars (PEG) [72] has been proposed [73, 74] as PEGs have a more formal basis compared to the diverging definitions and use of REGEXES.

While REGEXES are very powerful in many application scenarios such as log parsing [1, 2], their creation and maintenance are difficult and cumbersome [1, 2]. Thus researchers have previously investigated advanced methods for creating REGEXES. In the context of log parsing, synthesizing REGEXES from examples is a very interesting but non-trivial option [75]. Bartoli et al. [12] approached the REGEX synthesis problem by applying *multi-objective genetic programming* (GP), a paradigm rooted in evolutionary principles [76]. Later work [13] provides support for additional features, including alternatives and lookarounds. In a related study, Wang et al. [77] generated patterns for use in filtering tasks. Very recently, Chen et al. [78] introduced the concept of semantic regular expressions with a corresponding synthesis approach specifically targeted at data

extraction scenarios. Semantic REGEXES generalize standard REGEXES by considering a type and some additional criterion. For instance, the match needs to be a city (type) in Germany (criterion).

Further work focused on generating more robust REGEXES [18, 79]. Synthesizing REGEXES from natural language has been a promising research direction in recent years [43, 67, 80, 81]. Despite promising results, these approaches are not yet ready for practical use [16]. The benchmarks used for validation only contain short REGEXES and natural language descriptions with limited vocabulary. Real-world REGEXES are longer and more complicated, requiring more complex natural language descriptions.

A recent study by Zhang et al. [21] demonstrated that high-quality results can be achieved with current large language models (LLMs) even without model fine-tuning. The authors criticized the use of *sequence-to-sequence* models in previous works, which generate REGEX patterns character after character from left to right. It was recognized that this is not aligned with the order in which REGEXES are evaluated during pattern matching, which introduces the risk of syntactically incorrect patterns.

In contrast, their methodology is based on the concept of *chain-of-thought prompting* [67], which involves constructing multiple prompts in a step-by-step manner. This approach has been demonstrated to enhance the performance and interpretability of results produced by pre-trained LLMs. In their approach, the authors presented a novel REGEX formulation named *chain-of-inference*, where each chain represents a single sub-REGEX. During generation, these sub-problems are solved in the order of pattern matching, which mimics the human way of thinking. At the same time, no changes to the model or its training process are required.

To overcome some limitations of the approaches presented above, others have suggested multi-modal approaches [16, 17]. Here, a synthesis from natural language is the first step with a resulting incomplete pattern. In a second step, provided examples are used to complete the patterns. While promising, in those approaches errors propagate in a way that if the initially generated pattern contains errors, so will the final pattern. Li et al. [15] compensate for this by first generating complete patterns from natural language and then checking against provided examples and correcting if required. The described two-step approaches bear similarity to our work presented here as also our conversion requires multiple steps to ensure results with a high conversion rate.

As shown here, the conversion and generation of REGEXES has been intensively studied but to the best of our knowledge, no approach to the translation between pattern languages — as we have demonstrated — exists.

A key issue with REGEXES in real-world applications is an effect called *catastrophic backtracking* [14, 18, 19, 34, 35] which can occur when matchers with backtracking behavior are used (as it is common in many programming languages). This will cause super-linear runtime complexity for the matcher which is a functional issue but could also

lead to a vulnerability called Regular Expression Denial of Service (REDoS) [82]. Besides great care in REGEX design and thorough edge-case testing, approaches to combat this issue have been proposed, often by using static analysis of program code to detect vulnerable REGEXES [35, 83, 84] or by synthesizing safe ones [18]. Languages like DPL that do not strive for universal applicability but are targeted at specific use-cases do not employ backtracking exactly to avoid such issues. In usage scenarios such as the ones sketched in this work, converting REGEXES to such a representation has to be done in multiple steps to achieve a high degree of compatibility.

8. Conclusion

In this work, we presented REPTILE, an approach that combines a rule-based approach for converting REGEXES to DPL patterns with a best-effort approach for cases where a full conversion is not possible. Furthermore, we presented REPTILE's pattern optimization approach to optimize the obtained DPL patterns by predicting high-level matchers. For that, we explored the capabilities of using GPT-4 and a zero-shot prompting strategy.

The evaluation of REPTILE's rule-based conversion with 946 REGEXES collected from a large company showed that REPTILE safely converted 73.7% of them. For the remaining 26.3% REGEXES it provided a best-effort conversion that required the input from the user to safely convert them. The evaluation of REPTILE's pattern optimization with 23 other real-world REGEXES collected from 13 different, popular technologies showed an average F1-score and MCC above 0.91 across five high-level DPL matchers. These results have ample practical implications for companies that migrate or migrated to a modern log analytics platform, such as Dynatrace. They can use REPTILE to automatically and safely convert their REGEXES to DPL patterns.

In future work, we will extend REPTILE to consider other source and target pattern languages, such as Grok. Furthermore, we plan to extend our evaluation to REGEXES from other business domains. Regarding REPTILE's pattern optimization, we plan to extend our approach to consider further high-level matchers, such as HEXINT or CREDITCARD. Furthermore, we plan to explore other LLMs and prompting strategies, such as *few-shot prompting* [45] or *chain-of-thought prompting* [67].

Acknowledgements

We would like to thank Dynatrace Austria GmbH for funding and supporting this project. Furthermore, we would like to thank the industrial partner of Dynatrace to provide the dataset of REGEXES used for investigating RQ1 and RQ2. We also acknowledge that some diagrams in this work contain icons sourced from *Flaticon*¹².

¹²<https://flaticon.com/>

References

- [1] T. Zhang, H. Qiu, G. Castellano, M. Rifai, C. S. Chen, F. Pianese, System Log Parsing: A Survey, *IEEE Transactions on Knowledge and Data Engineering* 35 (2023) 8596–8614.
- [2] P. He, J. Zhu, Z. Zheng, M. R. Lyu, Drain: An Online Log Parsing Approach with Fixed Depth Tree, in: 2017 IEEE International Conference on Web Services (ICWS), 2017, pp. 33–40. doi:10.1109/ICWS.2017.13.
- [3] Q. Fu, J. Zhu, W. Hu, J.-G. Lou, R. Ding, Q. Lin, D. Zhang, T. Xie, Where Do Developers Log? An Empirical Study on Logging Practices in Industry, in: Companion Proceedings of the 36th International Conference on Software Engineering, ICSE Companion 2014, Association for Computing Machinery, 2014, p. 24–33. URL: <https://doi.org/10.1145/2591062.2591175>. doi:10.1145/2591062.2591175.
- [4] B. Debnath, M. Solaimani, M. A. G. Gulzar, N. Arora, C. Lumezanu, J. Xu, B. Zong, H. Zhang, G. Jiang, L. Khan, LogLens: A Real-Time Log Analysis System, in: 2018 IEEE 38th international conference on distributed computing systems (ICDCS), IEEE, 2018, pp. 1052–1062.
- [5] V.-H. Le, H. Zhang, Log Parsing with Prompt-based Few-shot Learning, in: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), 2023, pp. 2438–2449. doi:10.1109/ICSE48619.2023.00204.
- [6] H. Mi, H. Wang, Y. Zhou, M. R.-T. Lyu, H. Cai, Toward Fine-Grained, Unsupervised, Scalable Performance Diagnosis for Production Cloud Computing Systems, *IEEE Transactions on Parallel and Distributed Systems* 24 (2013) 1245–1255.
- [7] R. Vaarandi, M. Pihelgas, Using Security Logs for Collecting and Reporting Technical Security Metrics, in: 2014 IEEE Military Communications Conference, 2014, pp. 294–299. doi:10.1109/MILCOM.2014.53.
- [8] H. Hamooni, B. Debnath, J. Xu, H. Zhang, G. Jiang, A. Mueen, LogMine: Fast Pattern Recognition for Log Analytics, in: Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 1573–1582.
- [9] Splunk Inc., About Splunk regular expressions, <https://docs.splunk.com/Documentation/SCS/current/Search/AboutSplunkregularexpressions>, 2023. [Online; accessed 2024-02-22].
- [10] Z. Zhong, J. Guo, W. Yang, T. Xie, J.-G. Lou, T. Liu, D. Zhang, Generating Regular Expressions from Natural Language Specifications: Are We There Yet?, in: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [11] A. Bartoli, G. Davanzo, A. De Lorenzo, M. Mauri, E. Medvet, E. Sorio, Automatic Generation of Regular Expressions from Examples with Genetic Programming, in: Proceedings of the 14th annual conference companion on Genetic and evolutionary computation, 2012, pp. 1477–1478.
- [12] A. Bartoli, G. Davanzo, A. De Lorenzo, E. Medvet, E. Sorio, Automatic Synthesis of Regular Expressions from Examples, *Computer* 47 (2014) 72–80.
- [13] A. Bartoli, A. De Lorenzo, E. Medvet, F. Tarlao, Inference of Regular Expressions for Text Extraction from Examples, *IEEE Transactions on Knowledge and Data Engineering* 28 (2016) 1217–1230.
- [14] J. E. F. Friedl, *Mastering Regular Expressions*, 3rd ed., O'Reilly Media, Inc., 2006.
- [15] Y. Li, S. Li, Z. Xu, J. Cao, Z. Chen, Y. Hu, H. Chen, S.-C. Cheung, TransRegex: Multi-modal Regular Expression Synthesis by Generate-and-Repair, in: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), IEEE, 2021, pp. 1210–1222.
- [16] X. Ye, Q. Chen, X. Wang, I. Dillig, G. Durrett, Sketch-Driven Regular Expression Generation from Natural Language and Examples, *Transactions of the Association for Computational Linguistics* 8 (2020) 679–694.
- [17] Q. Chen, X. Wang, X. Ye, G. Durrett, I. Dillig, Multi-modal Synthesis of Regular Expressions, in: Proceedings of the 41st ACM SIGPLAN conference on programming language design and implementation, 2020, pp. 487–502.

- [18] Y. Li, Z. Xu, J. Cao, H. Chen, T. Ge, S.-C. Cheung, H. Zhao, FlashRegex: Deducing Anti-ReDoS Regexes from Examples, in: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, 2020, pp. 659–671.
- [19] J. C. Davis, L. G. Michael IV, C. A. Coghlan, F. Servant, D. Lee, Why Aren't Regular Expressions a Lingua Franca? An Empirical Study on the Re-use and Portability of Regular Expressions, in: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 443–454.
- [20] L. G. Michael, J. Donohue, J. C. Davis, D. Lee, F. Servant, Regexes are Hard: Decision-Making, Difficulties, and Risks in Programming Regular Expressions, in: 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2019, pp. 415–426. doi:10.1109/ASE.2019.00047.
- [21] S. Zhang, X. Gu, Y. Chen, B. Shen, InfeRE: Step-by-Step Regex Generation via Chain of Inference, in: 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2023, pp. 1505–1515. doi:10.1109/ASE56229.2023.00111.
- [22] C. Chapman, P. Wang, K. T. Stolee, Exploring Regular Expression Comprehension, in: 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE), 2017, pp. 405–416. doi:10.1109/ASE.2017.8115653.
- [23] Elasticsearch B.V., Grok processor, <https://www.elastic.co/guide/en/elasticsearch/reference/current/grok-processor.html>, 2024. [Online; accessed 2024-02-22].
- [24] Datadog Inc., Parsing, https://docs.datadoghq.com/logs/log_configuration/parsing/, 2024. [Online; accessed 2024-02-22].
- [25] Dynatrace LLC., Dynatrace Pattern Language, <https://docs.dynatrace.com/docs/platform/grail/dynatrace-pattern-language>, 2024. [Online; accessed 2024-02-22].
- [26] Dynatrace LLC., Extraction and parsing commands, <https://docs.dynatrace.com/docs/platform/grail/dynatrace-query-language/commands/extraction-and-parsing-commands>, 2024. [Online; accessed 2024-02-22].
- [27] G. Siegfried, P. Byrne, M. Bangera, M. Crossley, 2024 Gartner® Magic Quadrant™ for Observability Platforms, Technical Report, Gartner, Inc., 2024.
- [28] M. Fitzgerald, Introducing Regular Expressions, O'Reilly Media, Inc., 2012.
- [29] S. C. Kleene, Representation of Events in Nerve Nets and Finite Automata, Princeton University Press, 1956, pp. 3–42. doi:10.1515/9781400882618-002.
- [30] N. Chomsky, Three Models for the Description of Language, IRE Transactions on Information Theory 2 (1956) 113–124.
- [31] P. Linz, S. H. Rodger, An Introduction to Formal Languages and Automata, Jones & Bartlett Learning, 2022.
- [32] D. Moseley, M. Nishio, J. Perez Rodriguez, O. Saarikivi, S. Toub, M. Veanes, T. Wan, E. Xu, Derivative Based Nonbacktracking Real-World Regex Matching with Backtracking Semantics, Proceedings of the ACM on Programming Languages 7 (2023) 1026–1049.
- [33] P. Hazel, PCRE - Perl Compatible Regular Expressions, <https://pcre.org/>, 1997. [Online; accessed 2024-07-10].
- [34] H. Fujinami, I. Hasuo, Efficient Matching with Memoization for Regexes with Look-around and Atomic Grouping, in: European Symposium on Programming, Springer, 2024, pp. 90–118.
- [35] M. Berglund, F. Drewes, B. van der Merwe, Analyzing catastrophic backtracking behavior in practical regular expression matching, Electronic Proceedings in Theoretical Computer Science 151 (2014) 109–123.
- [36] Dynatrace LLC., Log processing grammar, <https://docs.dynatrace.com/docs/platform/grail/dynatrace-pattern-language/log-processing-grammar>, 2024. [Online; accessed 2024-09-14].
- [37] Splunk Inc., About the search language, <https://docs.splunk.com/Documentation/SplunkCloud/latest/Search/Aboutthesearchlanguage>, 2017. [Online; accessed 2024-03-29].
- [38] Splunk Inc., rex, <https://docs.splunk.com/Documentation/SplunkCloud/latest/SearchReference/Rex>, 2023. [Online; accessed 2024-03-29].
- [39] The regexpp authors, regexpp, <https://github.com/mysticatea/regexpp>, 2021. [Online; accessed 2024-03-29].
- [40] W. Su, R. Li, C. Peng, H. Chen, Algorithms for Checking Intersection Non-emptiness of Regular Expressions, in: Theoretical Aspects of Computing – ICTAC 2023, Springer Nature Switzerland, 2023, pp. 216–235.
- [41] The greenery authors, greenery, <https://github.com/qntm/greenery>, 2024. [Online; accessed 2024-04-19].
- [42] M. Almeida, N. Moreira, R. Reis, Testing the Equivalence of Regular Languages, in: Electronic Proceedings in Theoretical Computer Science, volume 3, 2009, pp. 47–57. doi:10.4204/EPTCS.3.4.
- [43] Z. Zhong, J. Guo, W. Yang, J. Peng, T. Xie, J.-G. Lou, T. Liu, D. Zhang, SemRegex: A Semantics-Based Approach for Generating Regular Expressions from Natural Language Specifications, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 1608–1618. URL: <https://aclanthology.org/D18-1189>. doi:10.18653/v1/D18-1189.
- [44] OpenAI, GPT-4 Technical Report, arXiv preprint arXiv:2303.08774 (2023).
- [45] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, Advances in neural information processing systems 33 (2020) 1877–1901.
- [46] J. Fragner, log-regex, <https://github.com/fragjulian/log-regex>, 2024. [Online; accessed 2024-06-17].
- [47] Dynatrace LLC., Dynatrace App Toolkit, <https://developer.dynatrace.com/reference/app-toolkit/>, 2024. [Online; accessed 2024-04-07].
- [48] Dynatrace LLC., Strato design system, <https://developer.dynatrace.com/reference/design-system/>, 2024. [Online; accessed 2024-04-07].
- [49] Microsoft, Azure OpenAI Service models, <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models>, 2024. [Online; accessed 2024-05-18].
- [50] The reregexp authors, reregexp, <https://github.com/suchjs/reregexp>, 2018. [Online; accessed 2024-04-14].
- [51] Dynatrace LLC., DPL Architect, <https://docs.dynatrace.com/docs/platform/grail/dynatrace-pattern-language/dpl-architect>, 2024. [Online; accessed 2024-03-24].
- [52] The Regex101 community, Regex101 Community Patterns, <https://regex101.com/library>, 2024. [Online; accessed 2024-05-18].
- [53] ChaosSearch Inc., Regex Support, <https://docs.chaossearch.io/docs/regex-support>, 2024. [Online; accessed 2024-06-07].
- [54] A. A. Chuvakin, Public Security Log Sharing Site, <https://www.chuvakin.org/> and <https://log-sharing.dreamhosters.com/>, 2010. [Online; accessed 2024-05-18].
- [55] Elasticsearch B.V., Elastic examples, <https://github.com/elasticsearch/examples/tree/master>, 2023. [Online; accessed 2024-06-07].
- [56] J. Zhu, S. He, P. He, J. Liu, M. R. Lyu, Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics, in: 2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE), IEEE Computer Society, 2023, pp. 355–366. URL: <https://doi.ieeecomputersociety.org/10.1109/ISSRE59848.2023.00071>.
- [57] R. Srinivasan, Sample cloudfront access logs, <https://github.com/aws-samples/amazon-cloudfront-log-analysis>, 2018. [Online; accessed 2024-06-07].
- [58] Amazon Web Services Inc., Public DNS query logging, <https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/query-logs.html>, 2024. [Online; accessed 2024-06-07].
- [59] Amazon Web Services Inc., Amazon S3 server access log format, <https://docs.aws.amazon.com/AmazonS3/latest/userguide/LogFormat.html>, 2024. [Online; accessed 2024-06-07].
- [60] Amazon Web Services Inc., Flow log record examples, <https://docs.aws.amazon.com/vpc/latest/userguide/flow-logs-records-examples.html>, 2024. [Online; accessed 2024-06-07].

- [61] kyhau, VPC Flow Log examples, <https://kyhau.github.io/aws-notebook/VpcFlowLogs.html>, 2021. [Online; accessed 2024-06-07].
- [62] Pluralsight LLC., Working with AWS VPC Flow Logs for Network Monitoring, <https://www.pluralsight.com/cloud-guru/labs/aws/working-with-aws-vpc-flow-logs-for-network-monitoring>, 2021. [Online; accessed 2024-06-07].
- [63] D. Lentz, Key metrics for CoreDNS monitoring, <https://www.datadoghq.com/blog/coredns-metrics/>, 2023. [Online; accessed 2024-06-07].
- [64] DigitalOcean LLC., How to Customize CoreDNS for Kubernetes Clusters, <https://docs.digitalocean.com/products/kubernetes/how-to/customize-coredns/>, 2024. [Online; accessed 2024-06-07].
- [65] Coding Assistant, <https://prollm.toqan.ai/leaderboard/coding-assistant>, 2024. [Online; accessed 2024-10-25].
- [66] Microsoft, Customize a model with fine-tuning, <https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/fine-tuning>, 2024. [Online; accessed 2024-06-12].
- [67] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, *Advances in neural information processing systems* 35 (2022) 24824–24837.
- [68] V. M. Glushkov, The abstract theory of automata, *Russian Mathematical Surveys* 16 (1961) 1.
- [69] D. Ziadi, J.-M. Champarnaud, An optimal parallel algorithm to convert a regular expression into its glushkov automaton, *Theoretical computer science* 215 (1999) 69–87.
- [70] S. Bhargava, G. Purohit, Construction of a minimal deterministic finite automaton from a regular expression, *International Journal of Computer Applications* 15 (2011) 16–27.
- [71] A. Kumar, A. K. Verma, A novel algorithm for the conversion of parallel regular expressions to non-deterministic finite automata, *Applied Mathematics & Information Sciences* 8 (2014) 95.
- [72] B. Ford, Parsing expression grammars: a recognition-based syntactic foundation, in: *Proceedings of the 31st ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, 2004, pp. 111–122.
- [73] M. Oikawa, R. Ierusalimschy, A. Moura, Converting regexes to parsing expression grammars, in: *Proceedings of the 14th Brazilian symposium on programming languages, SBLP*, volume 10, 2010.
- [74] S. Medeiros, F. Mascarenhas, R. Ierusalimschy, From regexes to parsing expression grammars, *Science of Computer Programming* 93 (2014) 3–18.
- [75] S. Tariq, T. A. Rana, Automatic regex synthesis methods for english: a comparative analysis, *Knowledge and Information Systems* (2024) 1–31.
- [76] J. R. Koza, Genetic Programming as a Means for Programming Computers by Natural Selection, *Statistics and computing* 4 (1994) 87–112.
- [77] X. Wang, S. Gulwani, R. Singh, FIDEX: Filtering Spreadsheet Data using Examples, *ACM SIGPLAN Notices* 51 (2016) 195–213.
- [78] Q. Chen, A. Banerjee, Ç. Demiralp, G. Durrett, I. Dillig, Data extraction via semantic regular expression synthesis, *Proceedings of the ACM on Programming Languages* 7 (2023) 1848–1877.
- [79] M. Lee, S. So, H. Oh, Synthesizing Regular Expressions from Examples for Introductory Automata Assignments, in: *Proceedings of the 2016 ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences*, 2016, pp. 70–80.
- [80] J.-U. Park, S.-K. Ko, M. Coggnetta, Y.-S. Han, SoftRegex: Generating Regex from Natural Language Descriptions using Softened Regex Equivalence, in: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 6425–6431.
- [81] C. Hahn, F. Schmitt, J. J. Tillman, N. Metzger, J. Siber, B. Finkbeiner, Formal Specifications from Natural Language, *arXiv preprint arXiv:2206.01962* (2022).
- [82] S. Crosby, Denial of service through regular expressions, *USENIX Association*, Washington, D.C., 2003.
- [83] J. Kirrage, A. Rathnayake, H. Thielecke, Static analysis for regular expression denial-of-service attacks, in: J. Lopez, X. Huang, R. Sandhu (Eds.), *Network and System Security*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 135–148.
- [84] V. Wüstholtz, O. Olivo, M. J. Heule, I. Dillig, Static detection of dos vulnerabilities in programs that use regular expressions, in: *Tools and Algorithms for the Construction and Analysis of Systems: 23rd International Conference, TACAS 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22-29, 2017, Proceedings, Part II* 23, Springer, 2017, pp. 3–20.