Chapter 3

Decidable Languages

We give basic definitions for languages, operators on languages and, finally, regular languages. We provide the corresponding formalizations from our development and prove their correctness.

3.1 Definitions

We closely follow the definitions from [5]. An **alphabet** Σ is a finite set of symbols. A **word** w is a finite sequence of symbols chosen from some alphabet. We use $|w|$ to denote the **length** of a word w. ε denotes the empty word. Given two words $w_1 = a_1 \cdots a_n$ and $w_2 = b_1 \cdots b_m$, the **concatenation** of w_1 and w_2 is defined as $a_1 \cdots a_n \cdot b_1 \cdots b_m$ and denoted $w_1 \cdot w_2$ or just $w_1 w_2$. A **language** is a set of words. The **residual language** of a language L with respect to symbol a is the set of words u such that av is in L. The residual is denoted $res_a(L)$. We define Σ^k to be the **set of words of length k**. The **set of all words** over an alphabet Σ is denoted Σ^*, i.e.,$\Sigma^* = \bigcup_{k \in \mathbb{N}} \Sigma^k$. A language L is **decidable** if and only if there exists a boolean predicate that decides membership in L. We will only deal with decidable languages from here on. Throughout the remaining document, we will assume a fixed alphabet Σ.

We employ finite types to formalize alphabets. In the most definitions, alphabets will not be made explicit. However, the same name and type will be used throughout the entire development. Words are formalized as sequences over the alphabet. Decidable languages are represented by functions from *word* to *bool*.

```
Variable char: finType.
Definition word := seq char.
Definition language := pred word.
Definition residual x L : language := [preim cons x of L].
```

3.1.1 Operations on languages

The **concatenation** of two languages L_1 and L_2 is denoted $L_1 \cdot L_2$ and is defined as the set of words $w = w_1 w_2$ such that w_1 is in L_1 and w_2 is in L_2. The **Kleene closure** of a language L (also called Kleene star) is denoted L^* and is defined as the set of words $w = w_1 \cdots w_k$ such that $w_1 \ldots w_k$ are in L. ε is contained in L^* ($k = 0$). We define the **complement** of a language L as $L \backslash \Sigma^*$, which we write as $\neg L$. Furthermore, we make use of the standard set operations **union** and **intersection**.

For our Coq development, take Coquand and Siles's [2] implementation of these operators. plus and prod refer to union and intersection, respectively. Additionally, we also introduce the singleton languages (atom), the empty language (void) and the language containing only the empty word (eps).

```
Definition conc L1 L2 : language :=
  fun v => existsb i : 'I_(size v).+1, L1 (take i v) && L2 (drop i v).
Definition star L : language :=
  fix star v := if v is x :: v' then conc (residual x L) star v' else true.
Definition compl L : language := predC L.
Definition plus L1 L2 : language := [predU L1 & L2].
Definition prod L1 L2 : language := [predI L1 & L2].
Definition atom x : language := pred1 [:: x].
Definition void : language := pred0.
Definition eps : language := pred1 [::].
```

The definition of conc is based on a characteristic property of the concatenation of two languages. The following lemma proves this property.

Lemma 3.1.1. *Let $L_1, L_2, w = a_1 \cdots a_k$ be given. We have that*

$$w \in L_1 \cdot L_2 \iff \exists n \in \mathbb{N}. 0 < n \leq k \wedge a_1 \cdots a_{n-1} \in L_1 \wedge a_n \cdots a_k \in L_2.$$

Proof. "$\Rightarrow$" From $w \in L_1 \cdot L_2$ we have w_1, w_2 such that $w = w_1 w_2 \wedge w_1 \in L_1 \wedge w_2 \in L_2$. We choose $n := |w_1| + 1$. We then have that $a_1 \cdots a_{n-1} = a_1 \cdots a_{|w_1|} = w_1$ and $w_1 \in L_1$ by assumption. Similarly, $a_n \cdots a_k = a_{|w_1|+1} \cdots a_k = w_2$ and $w_2 \in L_2$ by assumption.
"$\Leftarrow$" We choose $w_1 := a_1 \cdots a_{n-1}$ and $w_2 := a_n \cdots a_k$. By assumption we have that $w = w_1 w_2$. We also have that $a_1 \cdots a_{n-1} \in L_1$ and $a_n \cdots a_k \in L_2$. It follows that $w_1 \in L_1$ and $w_2 \in L_2$. □

Listing 3.1: Formalization of lemma 3.1.1

```
Lemma concP : forall {L1 L2 v},
  reflect (exists2 v1, v1 \in L1 & exists2 v2, v2 \in L2 & v = v1 ++ v2)
          (v \in conc L1 L2).
```

The implementation of star makes use of a characteristic property of the Kleene closure, which is that any nonempty word in L^* can be seen as the concatenation of a nonempty word in L and a (possibly empty) word in L^*. The following lemma proves the correctness of this property.

Lemma 3.1.2. *Let $L, w = a_1 \cdots a_k$ be given. We have that*

$$w \in L^* \iff \begin{cases} a_2 \cdots a_k \in res_{a_1}(L) \cdot L^*, & \textit{if } |w| > 0; \\ w = \varepsilon, & \textit{otherwise.} \end{cases}$$

Proof. "⇒" We do a case distinction on $|w| = 0$.

1. $|w| = 0$. It follows that $w = \varepsilon$.

2. $|W| \neq 0$, i.e. $|w| > 0$. From $w \in L^*$ we have $w = w_1 \cdots w_l$ such that $w_1 \cdots w_l$ are in L. There exists a minimal n such that $|w_n| > 0$ and for all $m < n$, $|w_m| = 0$. Let $w_n = b_1 \cdots b_p$. We have that $b_2 \cdots a_p \in res_{b_1}(L)$. Furthermore, we have that $w_{n+1} \cdots w_l \in L^*$. We also have $a_1 = b_1$ and $w = a_1 \cdots a_k = w_n \cdots w_l$. Therefore, we have $a_2 \cdots a_k \in res_{a_1}(L) \cdot L^*$.

"⇐" We do a case distinction on the disjunction.

1. $w = \varepsilon$. Then $w \in L^*$ by definition.

2. $a_2 \cdots a_k \in res_{a_1}(L) \cdot L^*$. By lemma 3.1.1 we have n such that $a_2 \cdots a_{n-1} \in res_{a_1}(L)$ and $a_n \cdots a_k \in L^*$. By definition of res, we have $a_1 \cdots a_{n-1} \in L$. Furthermore, we also have $a_n \cdots a_k = w_1 \cdots w_l$ such that $w_1 \dots w_l$ are in L. We choose $w_0 := a_1 \cdots a_{n-1}$. It follows that $w = w_0 w_1 \cdots w_l$ with $w_0, w_1, \cdots w_l$ in L. Therefore, $w \in L^*$.

□

The formalization of lemma 3.1.2 connects the characteristic property used to define star to the mathematical definition. The propositional formula given here appears slightly more restrictive than our mathemitcal definition as it requires all words from L to be nonempty. Mathematically, however, this is no restriction. The advantage of this formalization is can be seen in later proofs, where induction over the length of a word $w \in L^*$ is necessary. In such places, we will use lemma 3.1.2 to get $w_1 \dots w_k \in L$ such that $w = w_1 \cdots w_k$ and proceed by induction over the list of words $w_1 \dots w_k$. Since every word w_i has a length of at least one, we shorten w by at least one character. Thus, we can use the inductive hypothesis of the induction over $|w|$ to prove the claim.

Listing 3.2: Formalization of lemma 3.1.2

```
Lemma starP : forall {L v},
  reflect (exists2 vv, all [predD L & eps] vv & v = flatten vv)
          (v \in star L).
```

Theorem 3.1.1. *The decidable languages are closed under concatenation, Kleene star, union, intersection and complement.*

Proof. We have already given algorithms for every operator. It remains to show that they are correct. For concatenation and the Kleene star, we have shown in lemma 3.1.1 and lemma 3.1.2 that the formalizations are equivalent to the mathematical definitions. The remaining operators (union, intersection, complement) can be applied directly to the result of the languages' boolean decision functions. □

3.2 Regular Languages

Definition 3.2.1. *The set of regular languages REG is defined to be exactly those languages generated by the following inductive definition:*

$$\frac{}{\emptyset \in \mathit{REG}} \qquad \frac{}{\{\varepsilon\} \in \mathit{REG}} \qquad \frac{a \in \Sigma}{\{a\} \in \mathit{REG}} \qquad \frac{L \in \mathit{REG}}{L^* \in \mathit{REG}}$$

$$\frac{L_1 \in \mathit{REG} \qquad L_2 \in \mathit{REG}}{L_1 \cup L_2 \in \mathit{REG}} \qquad \frac{L_1 \in \mathit{REG} \qquad L_2 \in \mathit{REG}}{L_1 \cdot L_2 \in \mathit{REG}}$$

3.2.1 Regular Expressions

Regular expressions mirror the definition of regular languages very closely. We will consider **extended regular expressions** that include negation (Not), intersection (And) and a single-symbol wildcard (Dot). Therefore, we have the following syntax for regular expressions:

$$r, s := \emptyset \mid \varepsilon \mid . \mid a \mid r^* \mid r + s \mid r \& s \mid rs \mid \neg r$$

The semantics of these constructors are as follows:

1. $\mathcal{L}(\emptyset) = \emptyset$

2. $\mathcal{L}(\varepsilon) = \{\varepsilon\}$

3. $\mathcal{L}(.) = \Sigma$

4. $\mathcal{L}(a) = \{a\}$

5. $\mathcal{L}(r^*) = \mathcal{L}(r)^*$

6. $\mathcal{L}(r + s) = \mathcal{L}(r) \cup \mathcal{L}(s)$

7. $\mathcal{L}(r\&s) = \mathcal{L}(r) \cap \mathcal{L}(s)$
8. $\mathcal{L}(rs) = \mathcal{L}(r) \cdot \mathcal{L}(s)$

Definition 3.2.2. *We say that two regular expressions r and s are equivalent if and only if*

$$\mathcal{L}(r) = \mathcal{L}(s).$$

We will later show that equivalence of regular expressions is decidable.

We take the implementation of regular expressions from Coquand and Siles's development ([2]), which is also based on SSReflect and comes with helpful infrastructure for our proofs. They also give a decision function the language associated with a regular expression.

Listing 3.3: Regular Expressions

```
Inductive regular_expression :=
 | Void
 | Eps
 | Dot
 | Atom of symbol
 | Star of regular_expression
 | Plus of regular_expression & regular_expression
 | And of regular_expression & regular_expression
 | Conc of regular_expression & regular_expression
 | Not of regular_expression .
Fixpoint mem_reg e :=
  match e with
  | Void => void
  | Eps => eps
  | Dot => dot
  | Atom x => atom x
  | Star e1 => star (mem_reg e1)
  | Plus e1 e2 => plus (mem_reg e1) (mem_reg e2)
  | And e1 e2 => prod (mem_reg e1) (mem_reg e2)
  | Conc e1 e2 => conc (mem_reg e1) (mem_reg e2)
  | Not e1 => compl (mem_reg e1)
  end.
```

We will later prove that this definition is equivalent to the inductive definition of regular languages in 3.2.1. In order to do that, we introduce a predicate on regular expressions that distinguishes **standard regular expressions** from **extended regular expressions** (as introduced above). The grammar of standard regular expression is as follows:

$$r, s := \emptyset \mid \varepsilon \mid a \mid r^* \mid r + s \mid rs$$

Connect standard reg-exp to reg. languages