Words and Languages
(part II)

Prof. Dan A. Simovici

UMB
1 Languages

2 Induction on Words
The main objects of study of the theory of formal languages are languages, which are defined as sets of certain sequences of symbols.

**Definition**

Let $A$ be an alphabet. A **language over $A$** is a subset of $A^*$.  

In other words, a language over $A$ is any set of words over this alphabet. For instance, $\{a, ab, abba\}$ is a finite language over the alphabet $\{a, b\}$. Similarly, $L = \{a^n \mid n \in \mathbb{N}\}$ is an infinite language over the same alphabet.
By identifying words of length 1 with the symbols of $A$, the set $A$ itself is a language over $A$.

Other special languages over $A$:
- the empty language $\emptyset$,
- the full language $A^*$, and
- the null language $\{\lambda\}$.

Since $A^*$ is a countably infinite set, the set of languages over $A$, $\mathcal{P}(A^*)$ is not countable.
If $L$ is a language over an alphabet $A$ and $A \subseteq A'$, then $L$ is also a language over the alphabet $A'$. Therefore, if \( \{L_0, \ldots, L_{n-1}\} \) is a finite collection of languages over the alphabets \( \{A_0, \ldots, A_{n-1}\} \), respectively, then for $0 \leq i \leq n - 1$, each $L_i$ is a language over $A = \bigcup_{1 \leq i \leq n} A_i$.

We denote by $A_L$ the alphabet that consists of those symbols that occur in at least one word in $L$. If $L$ is a language over $A$, then $A_L \subseteq A$. 
Definition

A language $L$ is $\lambda$-free if $\lambda \not\in L$. 
The set of all prefixes of the words of a language $L$ is denoted by $\text{PREF}(L)$. Similarly, the sets of infixes and suffixes of the words of $L$ are denoted by $\text{INFIX}(L)$ and $\text{SUFF}(L)$, respectively. Note that $L \subseteq L'$ implies $\Omega(L) \subseteq \Omega(L')$, where $\Omega$ is any of $\text{PREF}$, $\text{SUFF}$, or $\text{INFIX}$. Also, $\text{INFIX}(L)$, $\text{PREF}(L)$, $\text{SUFF}(L)$ contain the null word and include $L$. 
The sets of proper prefixes, proper suffixes and proper infixes of a language $L$ are denoted by $\text{PREF}_{\text{pr}}(L)$, $\text{INFIX}_{\text{pr}}(L)$, and $\text{SUFF}_{\text{pr}}(L)$, respectively. Since languages are sets of words, we can apply to them set-theoretical operations such as union, intersection, difference, etc. If $L \subseteq A^*$, the complement of $L$ with respect to the alphabet $A$ is $\overline{L}_A = A^* - L$. If $A$ is understood from the context, we may denote the complement $\overline{L}_A$ simply by $\overline{L}$. 
Definition
The **product** of two languages $L$ and $K$ over an alphabet $A$ is the language $LK$ defined by

$$LK = \{xy \mid x \in L \text{ and } y \in K\}.$$
Definition

Let $L \subseteq A^*$ be a language over the alphabet $A$. The $n^{\text{th}}$ power of $L$ is the language $L^n$ given by

$$L^0 = \{\lambda\}$$
$$L^{n+1} = L^n L$$

for every language $L$ and natural number $n$.

Note that $L^1 = L$. In general, $L^n$ is the set of all words that can be written as products of $n$ words of $L$. For $n = 0$, we regard $\lambda$ as the product of zero words of $L$. 
Example

Let $L = \{ab, a\}$ be a language over the alphabet $A = \{a, b\}$. We have

\[
L^0 = \{\lambda\} \\
L^1 = \{ab, a\} \\
L^2 = \{abab, aba, aab, aa\} \\
\vdots
\]
Definition

Let $L$ be a language. The language $L^*$, the star closure or Kleene closure of $L$, is the set

$$L^* = \bigcup \{ L^n \mid n \in \mathbb{N} \}.$$

The language $L^+$, the positive closure of $L$, is the set of words

$$L^+ = \bigcup \{ L^n \mid n \in \mathbb{P} \}.$$
- $L^*$ is the set of all words that can be written as a product of zero or more words of $L$.
- $L^+$ is the set of all words that can be written as a product of one or more words of $L$.
- Since $L^*$ includes the product of zero words of $L$, the null word $\lambda$ is a member of $L^*$ for any language $L$.
- $L \subseteq L^+ \subseteq L^*$ and $LL^* = L^*L = L^+$. Furthermore, if $u, v \in L^*$, then $uv \in L^*$. Also, note that $\lambda \in L^+$ if and only if $\lambda \in L$. 
Example

Let $L = \{a, bab\}$ be a language over the alphabet $A = \{a, b\}$. $L^*$ comprises the words $\lambda, a, bab, abab, baba, babbab, aa$, etc., and $L^+$ consists of the same words except for $\lambda$. 
We have the following properties for any language $L$:

\[
\begin{align*}
L^*L^* & = L^*, \\
L^*L & = LL^*, \\
L^+L & = LL^+ \\
(L^*)^* & = L^*, \\
(L^+)^+ & = L^+.
\end{align*}
\]

Also, note that $L \subseteq H$ implies $L^* \subseteq H^*$. 
Theorem

Let $A$ be an alphabet. We have:

1. $L_0 \cup (L_1 \cup L_2) = (L_0 \cup L_1) \cup L_2,$
2. $L_0(L_1L_2) = (L_0L_1)L_2,$
3. $L_0 \cup L_1 = L_1 \cup L_0,$
4. $L_0(L_1 \cup L_2) = (L_0L_1) \cup (L_0L_2),$ 
5. $(L_0 \cup L_1)L_2 = (L_0L_2) \cup (L_1L_2),$
6. $L \cup L = L,$

for every $L, L_0, L_1, L_2 \in \mathcal{P}(A^*).$
Theorem

For every language $L$ we have:

1. $\{\lambda\}L = L\{\lambda\} = L$,
2. $\emptyset L = L\emptyset = \emptyset$,
3. $L \cup \emptyset = \emptyset \cup L = L$,
4. $L^* = \{\lambda\} \cup L^* L$,
5. $L^* = (\{\lambda\} \cup L)^*$,
6. $\emptyset^* = \{\lambda\}$,
Theorem

Let $A$ be an alphabet and let $L$ be a language over $A$. We have

$$L^* = \{\lambda\} \cup L \cup L^2 \cup \cdots \cup L^k \cup L^{k+1} L^*,$$

for every $k \in \mathbb{N}$. 
Proof

It is clear that

\[ \{\lambda\} \cup L \cup L^2 \cup \ldots \cup L^k \cup L^{k+1} L^* \subseteq L^*, \]

for every \( k \in \mathbb{N} \).

Conversely, let \( x \in L^* \). We have either \( x = \lambda \) or \( x \in L^n \) for some \( n \geq 1 \). If \( n \leq k \), then \( x \in \{\lambda\} \cup L \cup L^2 \cup \ldots \cup L^k \cup L^{k+1} L^* \). If \( n > k \), then \( L^n = L^{k+1} L^{n-(k+1)} \subseteq L^{k+1} L^* \), so again \( x \in \{\lambda\} \cup L \cup L^2 \cup \ldots \cup L^k \cup L^{k+1} L^* \). Thus,

\[ \{\lambda\} \cup L \cup L^2 \cup \ldots \cup L^k \cup L^{k+1} L^* \subseteq L^*. \]
Corollary

For every language \( L \) we have:

\[ L^* = \{ \lambda \} \cup LL^*. \]

Proof.

The equality of the corollary follows from Theorem ?? by taking \( k = 0 \). \( \square \)
Definition

The reversal of a language \( L \subseteq A^* \) is the language \( L^R \) given by

\[
L^R = \{ x^R \mid x \in L \}.
\]

It is easy to see that \( (L^R)^R = L \) for every language \( L \).
Definition

Let $L, K$ be two languages over the alphabet $A$. The right quotient $LK^{-1}$ and the left quotient $K^{-1}L$ are the languages:

$$LK^{-1} = \{ x \in A^* \mid xy \in L \text{ for some } y \in K \}$$
$$K^{-1}L = \{ x \in A^* \mid yx \in L \text{ for some } y \in K \}.$$
Example

Let $A = \{a, b, c\}$ be an alphabet and $L = \{\lambda, a, ab, abc\}$ be a language over $A$. Consider the languages $K_0 = \{c\}$, $K_1 = \{b, c\}$, and $K_2 = \{b, c\}^*$ over the same alphabet. Then, we have

$$LK_0^{-1} = \{ab\},$$
$$LK_1^{-1} = \{a, ab\},$$
$$LK_2^{-1} = \{\lambda, a, ab, abc\}.$$
The left quotient of two languages can be expressed through the right quotient of related languages by the equality

$$K^{-1}L = \left( L^R (K^R)^{-1} \right)^R $$

and

$$LK^{-1} = \left( (K^R)^{-1} L^R \right)^R .$$
Consider the following equivalent statements.

1. $x \in K^{-1}L$;
2. $yx \in L$ for some $y \in K$;
3. $x^Rz \in L^R$ for some $z \in K^R$;
4. $x^R \in L^R(K^R)^{-1}$;
5. $x \in (L^R(K^R)^{-1})^R$. 
Example

Let $L$ be a language over an alphabet $A$. It is easy to see that the set $\text{PREF}(L)$ of prefixes of a language $L$ is $L(A^*)^{-1}$, while the set $\text{SUFF}(L)$ of suffixes of $L$ is $(A^*)^{-1}L$. 
Theorem

Let $L_0, L_1, K$ be languages over the alphabet $A$. We have

\[
(L_0 \cup L_1)^{-1} = L_0^{-1} \cup L_1^{-1}
\]

\[
(L_0 \cup L_1)^{-1} K = L_0^{-1} K \cup L_1^{-1} K
\]

\[
(L_0 \cap L_1)^{-1} \subseteq L_0^{-1} \cap L_1^{-1}
\]

\[
(L_0 \cap L_1)^{-1} K \subseteq L_0^{-1} K \cap L_1^{-1} K
\]

\[
L_0^{-1} K \cap L_1^{-1} \subseteq (L_0 - L_1)^{-1}
\]

\[
K^{-1}(L_0 \cup L_1) = K^{-1}L_0 \cup K^{-1}L_1
\]

\[
K^{-1}(L_0 \cap L_1) \subseteq K^{-1}L_0 \cap K^{-1}L_1
\]

\[
K^{-1}L_0 - K^{-1}L_1 \subseteq K^{-1}(L_0 - L_1).
\]
Theorem

For the languages $L, L_0, L_1 \subseteq A^*$ and $a \in A$ we have:

$$\{a\}^{-1}(L_0L_1) = \begin{cases} (\{a\}^{-1}L_0)L_1 & \text{if } \lambda \not\in L_0 \\ (\{a\}^{-1}L_0)L_1 \cup \{a\}^{-1}L_1 & \text{if } \lambda \in L_0 \end{cases}$$

$$\{a\}^{-1}L_1^* = (\{a\}^{-1}L_1)L_1^*.$$ 

Note that the first equality can also be written as:

$$\{a\}^{-1}(L_0L_1) = (\{a\}^{-1}L_0)L_1 \cup (\{\lambda\} \cap L_0)\{a\}^{-1}L_1.$$ 

The proof is a direct application of the definition.
If $K$ is a singleton, $K = \{u\}$, we denote the languages $\{u\}^{-1}L$ and $L\{u\}^{-1}$ by $u^{-1}L$ and $Lu^{-1}$, respectively. These languages are referred to as the left derivative of $L$ with respect to $u$ and the right derivative of $L$ with respect to $u$, respectively.
We have:

\[
\begin{align*}
(L_0 \cup L_1)u^{-1} &= L_0u^{-1} \cup L_1u^{-1} \\
(L_0 \cap L_1)u^{-1} &= L_0u^{-1} \cap L_1u^{-1} \\
L_0u^{-1} - L_1u^{-1} &= (L_0 - L_1)u^{-1} \\
u^{-1}(L_0 \cup L_1) &= u^{-1}L_0 \cup u^{-1}L_1 \\
u^{-1}(L_0 \cap L_1) &= u^{-1}L_0 \cap u^{-1}L_1 \\
u^{-1}L_0 - uv^{-1}L_1 &= u^{-1}(L_0 - L_1) \\
u^{-1}(v^{-1}L) &= (vu)^{-1}L \\
(Lu^{-1})v^{-1} &= L(vu)^{-1},
\end{align*}
\]

for all words \(u, v\).
Theorem

(Induction Principle for Words) Let $L \subseteq A^*$ be a set of words such that $\lambda \in L$, and $x \in L$ implies $xa \in L$ for every $a \in A$. Then, $L = A^*$. 
Example

Let $A$ be an alphabet, $x \in A^*$, and $a \in A$. We prove, by applying the Induction Principle for Words, that for every $x \in A^*$, if $xa = ax$, then $x = a^m$ for some $m \in \mathbb{N}$. Let

$$L = \{ x \in A^* \mid xa = ax \text{ implies } x = a^m \text{ for some } m \in \mathbb{N} \}.$$  

Since $\lambda a = a\lambda = a$ and $\lambda = a^0$, we have $\lambda \in L$. Suppose that $x \in L$ and consider the word $y = xa$. If $ya \neq ay$, then the implication in the definition of $L$ holds and $y \in L$. Therefore, assume that $ya = ay$. This implies $xaa = axa$, so $xa = ax$, which implies $x = a^m$ because we assumed $x \in L$. Thus, $y = xa = a^{m+1}$, so $y \in L$. By the Induction Principle for Words we have $L = A^*$. 