

언어 이론의 기초

언어는 문자들의 나열이다.

언어를 정의하려면 기본문자들의 집합(alphabet, vocabulary)를 먼저 정의하여야한다 (universe of discourse).

예) 이진수	$\{0, 1\}$
십진수	$\{0, 1, \dots, 9\}$
영어	$\{a, b, \dots, z\}$
한국어	$\{\text{ㄱ, ㄴ, \dots, ㅎ, ㅏ, ㅑ, \dots, ㅡ, ㅣ}\} + 1)\alpha$
컴퓨터 파일	Unicode

기본문자(문자, symbol)들의 집합을 Σ 라고하자. 언어는 기본문자 Σ 에서(over) 정의한다. 문자가 나열되면 문자열(string)이라 부른다. 이진수에서 01001, 1, 110들이 문자열의 예이다. 언어(language)는 문자열들의 집합으로 정의한다. 예를 들어 ㅎㅏㄱㅑㅓ 문자열은 한국어 어법에 맞는 한글 문자열이지만, 스ㅋㅊㄹ 은 한국어 어법에 맞는 한글 문자열이 **아니다**. 이 경우 한글 **어법에 맞는** 한국어 문자열만을 모은 집합을 한국어(언어; language)로 정의한다.

문자가 나열된 문자열의 들어있는 문자의 개수로 문자열의 길이를 정의할 수 있다. 예를 들어 $|01001| = 5$ 이고, $|1| = 1$, $|110| = 3$ 이다. 문자열의 **연결(concatenation)**을 정의할 수 있다. 예를 들어 문자열 school과 boy를 **연결**하면 schoolboy이다. 길이 6개짜리 문자열 school과 3개짜리 문자열 boy을 **연결**해서 길이 9개짜리 새로운 문자열 schoolboy를 만들었다. 길이가 0인 **빈 문자열(empty string)**을 생각하고 ϵ 으로 빈 문자열을 나타낸다고 하자. 빈 문자열은 어떤 문자열에 연결하여도 그 문자열을 바꾸지 않는다. 즉 ϵ 과 school을 연결하여도 school과 ϵ 을 연결하여도 그 결과는 그냥 school일 뿐이다. 이 특별한 문자열 ϵ 은 **연결연산에 항등연산자(identity element)**라고 부르고 $|\epsilon| = 0$ 이고 기본문자집합 Σ 가 무엇이던 항상 똑같이 ϵ 으로 정의된다.

언어의 전체집합(universe)은 무엇일까? 기본문자집합 $\Sigma = \{0, 1\}$ 인 이진수를 생각하자. 이진수 문자열은 길이가 1인 0, 1이 있고 길이 2인 00, 01, 10, 11과 길이 3인 000, 001, ..., 111등이 있다. 이것을 각각 $\Sigma^1, \Sigma^2, \Sigma^3$ 으로 정의한다면 이진수의 전집합은 $\Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$ 으로 표시할 수 있을 것이다. 여기에 빈 문자열 ϵ 을 가진 집합 $\{\epsilon\}$ 을 Σ^0 으로 정의하여 이진수의 전집합은 빈 문자열을 포함하여 $\Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$ 으로 정의할 수 있다.

(정의) 기본문자집합을 Σ 라고 하자. 기본문자집합의 **반복** $\Sigma^n (n \geq 0)$ 을 아래로 정의한다.

Basis	$\Sigma^0 = \{\epsilon\}$
Recursion	$\Sigma^n = \Sigma^{n-1}\Sigma$

(정의) 기본문자집합의 **반복합** Σ^+ 과 Σ^* 을 아래로 정의한다.

1) + α 에 관하여는 2장 참조

$$\Sigma^+ \stackrel{\text{def}}{=} \bigcup_{n \in N_1} \Sigma^n \stackrel{\text{def}}{=} \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots \quad \text{단 } N_1 = \{1, 2, 3, \dots\}.$$

$$\Sigma^* \stackrel{\text{def}}{=} \bigcup_{n \in N_0} \Sigma^n \stackrel{\text{def}}{=} \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \dots \quad \text{단 } N_0 = \{0, 1, 2, \dots\}.$$

$$\bigcup_{n \in N_1} \Sigma^n \stackrel{?}{=} \bigcup_{n=1}^{\infty} \Sigma^n.$$

$$\bigcup_{n \in N_1} \Sigma^n \stackrel{?}{=} \bigcup_{n=0}^{\infty} \Sigma^n.$$

네 가지 기본용어

언어는 기본문자집합 Σ 를 우선 정의하여야 한다. 이 때 (1) 문자 $a \in \Sigma$ 로 (2) 기본문자집합은 Σ 로, (3) 문자열 $x \in \Sigma^*$ 로 (4) 언어 $L \subseteq \Sigma^*$ 로 정의한다.

	원소	집합
길이 1	문자(symbol), $a \in \Sigma$	기본문자집합(vocabulary, alphabet), Σ
길이 0이상 ²⁾	문자열(string), $x \in \Sigma$	언어(language), $L \subseteq \Sigma^*$

기본문자³⁾ Σ 에서 정의한 언어 $L \subseteq \Sigma^*$ 에 대하여, 문자열 $x \in \Sigma^*$ 가 언어 L 에 속하면($x \in \Sigma^*$) 문장(sentence), 그렇지 않으면 ($x \notin \Sigma^*$) 비문장(non-sentence)이라 부른다.

Membership problem

언어 L 은 전집합 Σ^* 의 부분집합일 때 문자열 $x \in \Sigma^*$ 가 $x \in L$ 인가 $x \notin L$ 를 결정하는 문제를 언어 L 의 membership 문제라고 부른다.

$$L(x) = \text{true or false}$$

$$L: \Sigma^* \rightarrow \{\text{true}, \text{false}\}$$

Decision problem

집합 A 의 원소 $a \in A$ 에 대하여 특정 문제 f 의 결과가 긍정 또는 부정인 문제 f

$$f: A \rightarrow \{\text{true}, \text{false}\}$$

Membership problem은 decision problem의 일종이다.

2) 길이 0이상에는 길이 1도 포함되어 있다.
 3) 앞으로 기본문자는 기본문자집합과 같은 의미로 사용하겠다.