

언어는 어휘 위에서(over) 정의한다.

언어를 정의하려면 문자(symbol)들의 전집합 어휘(alphabet, vocabulary; 기본문자)를 먼저 정의하여야 한다. 문자(symbol)들의 집합 기본문자를 Σ^1 라고 하자.

예) 이진수 $\Sigma_{\text{이진수}} = \{0, 1\}$
 십진수 $\Sigma_{\text{십진수}} = \{0, 1, \dots, 9\}$
 영어 $\Sigma_{\text{영어}} = \{a, b, \dots, z\}$
 한글 $\Sigma_{\text{한글}} = \{\text{ㄱ}, \text{ㄴ}, \dots, \text{ㅎ}, \text{ㅏ}, \text{ㅑ}, \dots, \text{ㅣ}\} + \alpha^2$
 Text 파일 $\Sigma_{\text{Text파일}} = \text{unicode}$
 영어문장 $\Sigma_{\text{영어문장}} = \text{영어사전에 나온 단어}$

언어(language)는 어휘 Σ 위에서(over) 정의(universe of discourse)한다³⁾. 문자가 나열되면 문자열⁴⁾(string)이 된다. 어휘 $\{0, 1\}$ 에서 정의된 이진수 언어에서 1, 110, 00110등이 문자열의 예이다.

문자열 중에는 특정 언어(language)에 맞는 문자열이 있고 그렇지 않은 문자열이 있다. 그래서 언어(language)를 문자열의 집합으로 정의한다. 예를 들어 ㅎㅏㅑㅓㅕㅗ 문자열은 한글 어법에 맞는 한글 문자열이지만, ㅏㅑㅓㅕㅗ 문자열은 현대한글 어법에 맞지 않으므로 한글 문자열이 아니다. 이 경우 한글 어법에 맞는 한국어 문자열만을 모은 집합을 한글(언어; language)로 정의할 수 있다⁵⁾.

문자열의 길이는 그 문자열에 들어 있는 문자의 개수이다. 예를 들어 $|0| = 1$ 이고 $|110| = 3$, $|00110| = 5$ 이다. 문자열의 연결(concatenation, \cdot)을 정의할 수 있다. 예를 들어 문자열 school과 boy를 연결하면 $\text{school} \cdot \text{boy} = \text{schoolboy}^6)$ 이다. 길이 6개짜리 문자열 school과 길이 3개짜리 문자열 boy를 연결해서 길이 9개짜리 새로운 문자열 schoolboy를 만든다.

길이가 0인 빈 문자열(empty string)을 생각하고, $\epsilon^8)$ 이 빈 문자열을 나타낸다고 하자. 빈 문자열은 어떤 문자열과 연결하여도 그 문자열을 바꾸지 않는다. 즉 ϵ 과 school을 연결하거나 school과 ϵ 을 연결하여도, 결과는 그냥 school일 뿐이다⁹⁾. 빈 문자열 ϵ 을 연결 연산에서 항등원(identity element)¹⁰⁾이고 $|\epsilon| = 0$ 이고 기본문자 Σ 에 관계없이 항상 ϵ 으로 표시한다.

1) 그리스어 대문자 Sigma이다. 소문자는 σ 이다. 영어로는 S나 s에 대응한다.

2) $+\alpha$ 에 관하여는 2장 한글 모아쓰기 기계 참조

3) a language over Σ

4) 문자열 = a sequence of symbols. 고등학교 때 배운 수열은 숫자들의 나열임을 기억하자.

5) 2장 한글모아쓰기 기계 참조.

6) 이렇게 쓰는 것을 그대로쓰기(juxtaposed)라고 한다.

7) 연결기호 \cdot 는 생략하기도 한다(juxtaposed).

8) ϵ 은 그리스 문자 epsilon(e)이다. 어떤 교과서에서는 그리스 문자 Lambda인 $\Lambda(L)$ 나 $\lambda(l)$ 를 쓰기도 한다.

9) $\epsilon \cdot \text{school} = \text{school} \cdot \epsilon = \text{school}$.

10) $+$ 연산에서 항등원은 0이고, \times 연산에서 항등원은 1이다.

문자열의 전집합(universe)은 무엇일까?

어휘가 $\Sigma = \{0, 1\}$ 인 이진수 언어를 생각하자. 이진수 문자열은 길이가 1인 0, 1이 있고¹¹⁾ 길이가 2인 00, 01, 10, 11과 길이가 3인 000, 001, ..., 111등이 있다. 이것을 각각 Σ^1 , Σ^2 , Σ^3 으로 정의한다면 이진수의 전 집합은 $\Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$ 으로 표시할 수 있을 것이다. 여기에 길이가 0인 빈 문자열 ϵ 도 포함하여, $\{\epsilon\}$ 을 Σ^0 으로 정의하면, 이진수 전집합(universe)은 빈 문자열을 포함하여 $\Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$ 으로 정의할 수 있다.

(정의) 기본문자를 Σ 라고 하자. 기본문자 n 번 반복 $\Sigma^n (n \geq 0)$ 을 아래로 정의한다.

$$\text{Basis} \quad \Sigma^0 \stackrel{\text{B}}{=} \{\epsilon\}.$$

$$\text{Recursion} \quad \Sigma^n \stackrel{\text{R}}{=} \Sigma \cdot \Sigma^{n-1}, \text{ 단 } n \geq 1.$$

$$\begin{aligned} \text{(예)} \quad \{0, 1\}^2 &\stackrel{\text{R}}{=} \{0, 1\} \cdot \{0, 1\}^1 \stackrel{\text{R}}{=} \{0, 1\} \cdot \{0, 1\} \cdot \{0, 1\}^0 \stackrel{\text{B}}{=} \{0, 1\} \cdot \{0, 1\} \cdot \{\epsilon\} \\ &= \{0, 1\} \cdot \{0, 1\} = \{0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1\} = \{00, 01, 10, 11\}. \end{aligned}$$

(사실) $|\Sigma^n| = |\Sigma|^n (n \geq 0)$.

(증명) $n \geq 0$ 에 관한 수학적 귀납법(생략).

(정의) 기본문자 Σ 의 반복합 Σ^\dagger 과 Σ^* (¹²⁾)을 아래로 정의한다.

$$\Sigma^\dagger \stackrel{\text{B}}{=} \bigcup_{i \in N_1} \Sigma^i = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots \quad \text{단 } N_1 \stackrel{\text{B}}{=} \{1, 2, 3, \dots\}.$$

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

Σ^\dagger 는 길이가 1이상인 모든 문자열에 집합, Σ^* 은 빈 문자열 ϵ 도 포함하여 길이가 0이상인 모든 문자열의 집합을 나타내며, Σ^* 를 문자열의 전집합(universe)으로 본다.

$$\begin{aligned} \text{(질문)} \quad |\Sigma^*| &= |\Sigma|^0 + |\Sigma|^1 + |\Sigma|^2 + \dots + |\Sigma|^k + \dots \\ &= n^0 + n^1 + n^2 + \dots + n^k + \dots \quad (\text{단 } |\Sigma| = n \geq 2) \\ &= \lim_{k \rightarrow \infty} \frac{n^{k+1} - 1}{n - 1} = ? \infty. \end{aligned}$$

$$\text{(질문)} \quad |\Sigma^*| = ? |\Sigma^\dagger| + 1$$

(정리) $\Sigma^\dagger \subset \Sigma^*$ 이지만 $|\Sigma^*| = |\Sigma^\dagger| = \aleph$.

(증명) ???

연결(\cdot ; concatenation)에 정의역과 치역을 문자열($x \in \Sigma^*$)에서 언어($L \subseteq \Sigma^*$ 혹은 $L \in 2^{\Sigma^*}$)로 일반화한다.

(정의) $\cdot : 2^{\Sigma^*} \times 2^{\Sigma^*} \rightarrow 2^{\Sigma^*}$.

$$L, S \subseteq \Sigma^* \text{라 하자. } L \cdot S = \{x \cdot y \in \Sigma^* | x \in L, y \in S\}^{13}.$$

11) 문자는 길이가 1인 문자열의 일종이다.

12) Σ^\dagger 는 Σ dagger라고 읽고, Σ^* 는 Σ star라고 읽는다.

13) $L^* = L^\dagger \cup \{\epsilon\}$. If $\epsilon \notin L$, $L^\dagger = L^* - \{\epsilon\}$; otherwise $L^\dagger = L^*$.

네 가지 기본용어(terminologies)

언어는 기본문자 Σ 를 우선 정의하여야 한다. 이 때 언어이론에 네 가지 기본용어인 (1) 문자(symbol)는 $a \in \Sigma$ 로 (2) 어휘(vocabulary)는 Σ 로, (3) 문자열(string)은 $x \in \Sigma^*$ 로 (4) 언어(language)는 $L \subseteq \Sigma^*$ 또는 $L \in 2^{\Sigma^*}$ 로 쓴다.

	원소	집합
길이 1	문자 ($a \in \Sigma$)	어휘 (Σ)
길이 0이상 ¹⁴⁾	문자열 ($x \in \Sigma^*$)	언어 ($L \subseteq \Sigma^*$ 또는 $L \in 2^{\Sigma^*}$)

어휘 Σ 에서 정의한 언어 $L \subseteq \Sigma^*$ 에 대하여, 문자열 $x \in \Sigma^*$ 가 언어 L 에 속하면($x \in L$) 문장(sentence), 그렇지 않으면 ($x \notin L$) 문장이 아니(non-sentence: 비문)라고 한다.

원소문제(membership problem)

기본문자 Σ 에서 정의한 언어 $L \subseteq \Sigma^*$ 에 대하여, 임의의 문자열 $x \in \Sigma^*$ 가 문장인가 아닌가를 정하는 문제를 언어 L 의 원소문제라고 부른다.

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

$$L(x) = \mathbf{if } x \in L \rightarrow \mathbf{true} \mid x \notin L \rightarrow \mathbf{false} \mathbf{fi}.$$

결정문제(Decision problem, 문제)

집합 D 의 임의의 원소 $d \in D$ 에 대하여 $P(d)$ 의 결과가 **true** 또는 **false**인 함수 P 를 결정문제(문제)라 한다.

$$P: D \rightarrow \{\mathbf{true}, \mathbf{false}\}$$

언어의 원소문제는 정의역(domain) D 가 Σ^* 인 결정문제의 일종이다¹⁵⁾.

(정리) $|D| = \aleph$ 일 때, $|L| = |P| > \aleph$.

14) 문자는 길이가 1인 문자열이다.

15) 강의의 후반부에 가면, 언어와 문제를 구분하지 않을 것이다.