예) 이진수 $\Sigma_{이진수} = \{0, 1\}$
십진수 $\Sigma_{십진수} = \{0, 1, \cdots, 9\}$
영어 $\Sigma_{영어} = \{a, b, \cdots, z\}$
한글 $\Sigma_{한글} = \{ㄱ, ㄴ, \cdots, ㅎ, ㅏ, ㅑ, \cdots, ㅣ\} + \alpha$[1]
Text 파일 $\Sigma_{Text파일}$ = unicode
영어문장 $\Sigma_{영어문장}$ = 영어사전에 나온 단어와 구두점(, . ? !등)

언어(language)는 어휘 Σ 위에서(over) 정의[2](universe of discourse)한다. 문자가 나열되면 문자열[3](string)이 된다. 어휘 {0, 1}에서 정의된 이진수에서 1, 110, 00110등이 문자열의 예이다.

문자열 중에는 특정 언어(language)에 맞는 문자열이 있고 그렇지 않은 문자열이 있다. 그래서 언어(language)를 문자열의 **집합**으로 정의한다. 예를 들어 ㅎㅏㄱㄱㅛ 문자은 "학교"라는 한글 어법에 **맞는** 한글 문자열이지만, ㅅㅋㅜㄹ 문자열은 "쑬"은 현대한글 어법에 **맞지 않으므로** 한글 문자열이 아니다. 이 경우 한글 **어법에 맞는** 한국어 **문자열**만을 모은 **집합**을 한글(**언어**; **language**)로 정의할 수 있다[4].

문자열의 **길이**는 그 문자열에 들어 있는 문자의 개수이다. 예를 들어 |0| = 1이고 |110| = 3, |00110| = 5이다. 문자열의 **연결**(**concatenation**, ·)을 정의할 수 있다. 예를 들어 길이 6개짜리 문자열 school과 길이 3개짜리 boy를 **연결**하여 길이 9개짜리 새로운 문자열 school · boy = schoolboy[5]를 만든다.

길이가 0인 **빈 문자열**(**empty string**)을 생각하고, ϵ[6]이 **빈 문자열**을 나타낸다고 하자. **빈 문자열**은 어떤 문자열과 **연결**하여도 그 문자열을 바꾸지 않는다. 즉 ϵ과 school을 연결하거나 school과 ϵ을 연결하여도, 결과는 그냥 school일 뿐이다[7]. **빈 문자열** ϵ을 **연결 연산**에서 **항등원**(identity element)[8]이고 $|\epsilon| = 0$이고 기본문자 Σ에 관계없이 항상 ϵ으로 표시한다.

---

1) + α에 관하여는 2장 한글 모아쓰기 오토마타 참조
2) a language over Σ
3) 문자열 = a sequence of symbols. 고등학교 때 배운 수열은 수들의 나열임을 기억하자.
4) 2장 한글모아쓰기 오토마타 참조. 현대한글은 모두 $21 \times 19 \times 28 = 1{,}1172$자로 정의된다(유니코드).
5) 이렇게 ·을 생략하고 그대로 쓰는 것을 그대로쓰기(juxtaposed)라고 한다.
6) ϵ은 그리스 문자 epsilon(영어 e)이다. 어떤 교과서에서는 그리스 문자 Lambda인 Λ(영어 L)나 λ(영어 l)를 쓰기도 한다.
7) ϵ · school = school · ϵ = school.
8) +연산에서 항등원은 0이고, ×연산에서 항등원은 1이다.

KAIST 전산학과 최광무

문자열의 전집합(universe)은 무엇일까?

어휘가 $\Sigma = \{0, 1\}$인 이진수 언어를 생각하자. 이진수 문자열은 길이가 1인 0, 1이 있고[9] 길이가 2인 00, 01, 10, 11과 길이가 3인 000, 001, …, 111등이 있다. 이것을 각각 Σ^1, Σ^2, Σ^3으로 정의한다면 이진수의 전 집합은 $\Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \cdots$으로 표시할 수 있을 것이다. 여기에 길이가 0인 빈 문자열(empty string) ϵ도 **포함**하여, $\{\epsilon\}$을 Σ^0으로 정의하면, 이진수 전집합(universe)은 빈 문자열을 포함하여 $\Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \cdots$으로 정의할 수 있다.

(정의) 기본문자(symbol)의 집합(vocabulary)를 Σ라고 하자. 기본문자가 n번 **반복**된 문자열(string) $\Sigma^n (n \geq 0)$을 아래와 같이 정의한다.

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

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

(예) $\{0, 1\}^2 \stackrel{\text{def}}{=}_{\text{R}} \{0, 1\} \cdot \{0, 1\}^1 \stackrel{\text{def}}{=}_{\text{R}} \{0, 1\} \cdot \{0, 1\} \cdot \{0, 1\}^0 \stackrel{\text{def}}{=}_{\text{B}} \{0, 1\} \cdot \{0, 1\} \cdot \{\epsilon\}$
$= \{0, 1\} \cdot \{0, 1\} = \{0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1\} = \{00, 01, 10, 11\}.$

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

(증명) $n \geq 0$에 관한 수학적 귀납법(생략).

(정의) 기본문자 Σ의 **반복합** $\Sigma^\dagger$과 Σ^*[10]을 아래로 정의한다.

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

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

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

(질문) $|\Sigma^*| = |\Sigma|^0 + |\Sigma|^1 + |\Sigma|^2 + \cdots + |\Sigma|^k + \cdots$
$= n^0 + n^1 + n^2 + \cdots + n^k + \cdots \quad (\text{단 } |\Sigma| = n \geq 2)$
$= \lim_{k \to \infty} \frac{n^{k+1} - 1}{n - 1} \stackrel{?}{=} \infty.$

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

(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.($\cdot : \Sigma^* \times \Sigma^* \to \Sigma^*$)

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

$L, S \subseteq \Sigma^*$라 하자. $L \cdot S = \{x \cdot y \in \Sigma^* | x \in L, y \in S\}$[11].

---

9) 문자(symbol)는 길이가 1인 문자열(string)의 일종이다.

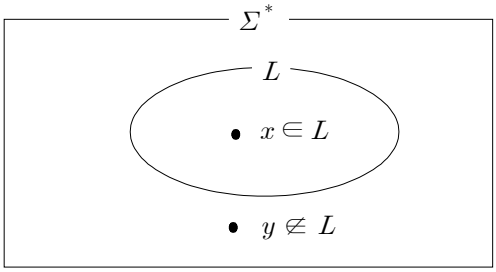
10) $\Sigma^\dagger$는 Σ dagger라고 읽고, Σ^*는 Σ star라고 읽는다.

네 가지 기본용어(terminologies)

언어는 기본문자 Σ를 우선 정의하여야 한다. 이 때 언어이론에 네 가지 기본용어인 (1) **문자(symbol)**는 $a \in \Sigma$로 (2) **어휘(vocabulary)**는 Σ로, (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** (symbols) $a, b, c, \cdots \in \Sigma$	**어휘, 기본문자** (vocabuary, alphabet) Σ
길이 0이상[12]	**문자열** (string) $x, y, z, \cdots \in \Sigma^*$	**언어** (language) $L, S, T, \cdots \subseteq \Sigma^*$ 또는 $L, S, T, \cdots \in 2^{\Sigma^*}$

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



원소문제(membership problem)

기본문자 Σ에서 정의한 언어 $L \subseteq \Sigma^*$에 대하여, 임의의 문자열 $x \in \Sigma^*$가 **문장인가 아닌가**를 정하는 문제를 언어 L의 원소문제라고 부른다.

$$L\colon \Sigma^* \to \{\mathbf{true}, \mathbf{false}\}$$
$$L(x) = \mathbf{if}\ x \in L \to \mathbf{true} \mid x \notin L \to \mathbf{false}\ \mathbf{fi}.$$

결정문제(Decision problem, 문제)

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

$$P\colon D \to \{\mathbf{true}, \mathbf{false}\}$$

언어의 원소문제는 정의역(domain) D가 Σ^*이고 함수 P가 L인 결정문제의 일종이다[13].

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

정의역이 countably infinite인 언어의 원소문제는 uncountable이다.

---

11) $L^* = L^\dagger \cup \{\epsilon\}$. *If* $\epsilon \notin L$, $L^\dagger = L^* - \{\epsilon\}$; *otherwise* $L^\dagger = L^*$.
12) 문자는 길이가 1인 문자열이다.
13) 강의의 후반부에 가면, **언어**와 **문제**를 구분하지 않을 것이다.