

一般言語の定義

水谷 正大

mizutani@ic.daito.ac.jp

2013 April

目次

- ① アルファベット
- ② 語と言語
- ③ 言語のファミリー
- ④ 言語の認識と生成
- ⑤ 言語機械

アルファベット

文字 $\in$ アルファベット

- 基本ラテン文字 (ローマ字) $\{a, b, c, \dots, z, A, B, \dots, Y, Z\}$ 52 字
- ひらがな $\{あ, い, う, え, お, \dots, ゐ, ゑ, を, ん\}$ (小学校令施行規則) 48 文字
- ASCII $\{a, b, \dots, y, z, A, B, \dots, Z, 0, 1, \dots, 9, \&, , \dots, \text{Del}\}$ 128 文字
- JIS X 0208 $\{\dots, \blacksquare, \dots, 大, \dots, 学, \dots, \alpha, \dots, \text{Я}, \dots\}$ 6,879 図形文字
- 漢字 $\{厶, 伊, 宇, 江, 尾, \dots, 鬱, \dots\}$ 今昔文字鏡 15 万字
- Unicode(世界で使われている文字を共通の文字集合とする方法の 1 つ)

文字の符号化

どんな文字集合も 2 つの記号で**符号化**可能

- bit(0 と 1) で符号化
- n bits で符号化すると 2^n 個からなる文字集合が表現可能
- 符号化された文字を数 (2 進数) と見なしてもよい

基本ラテン文字集合 52 文字

- 5 bits では足りない ($2^5 = 32$)
- 6 bits で表現できる ($2^6 = 64$)

7ビット符号化したASCII文字表

文字	コード 10進 16進	文字	コード 10進 16進	文字	コード 10進 16進	文字	コード 10進 16進	文字	コード 10進 16進	文字	コード 10進 16進	文字	コード 10進 16進
NUL	0 00	DLE	16 10	SP	32 20		0 48 30	@	64 40	P	80 50	`	96 60
SOH	1 01	DC1	17 11	!	33 21	1	49 31	A	65 41	Q	81 51	a	97 61
STX	2 02	DC2	18 12	"	34 22	2	50 32	B	66 42	R	82 52	b	98 62
ETX	3 03	DC3	19 13	#	35 23	3	51 33	C	67 43	S	83 53	c	99 63
EOT	4 04	DC4	20 14	\$	36 24	4	52 34	D	68 44	T	84 54	d	100 64
ENQ	5 05	NAK	21 15	%	37 25	5	53 35	E	69 45	U	85 55	e	101 65
ACK	6 06	SYN	22 16	&	38 26	6	54 36	F	70 46	V	86 56	f	102 66
BEL	7 07	ETB	23 17	'	39 27	7	55 37	G	71 47	W	87 57	g	103 67
BS	8 08	CAN	24 18	(	40 28	8	56 38	H	72 48	X	88 58	h	104 68
HT	9 09	EM	25 19	)	41 29	9	57 39	I	73 49	Y	89 59	i	105 69
NL*	10 0a	SUB	26 1a	*	42 2a	:	58 3a	J	74 4a	Z	90 5a	j	106 6a
VT	11 0b	ESC	27 1b	+	43 2b	;	59 3b	K	75 4b	[	91 5b	k	107 6b
NP	12 0c	FS	28 1c	,	44 2c	<	60 3c	L	76 4c	\	92 5c	l	108 6c
CR	13 0d	GS	29 1d	-	45 2d	=	61 3d	M	77 4d	]	93 5d	m	109 6d
SO	14 0e	RS	30 1e	.	46 2e	>	62 3e	N	78 4e	^	94 5e	n	110 6e
SI	15 0f	US	31 1f	/	47 2f	?	63 3f	O	79 4f	_	95 5f	o	111 6f

2⁷ = 128 文字 *NL: "New Line" (改行)の略。"LF" (Line Feed)と呼ばれることもある

最初に制定された日本語文字集合 JIS X0201

ASCII と比較すると

- 92(0x5C) 番：バックスラッシュ ('\\') ⇒ 円記号 ('¥')
- 126(0x7E) 番：チルダ ('~') ⇒ オーバーライン ('_')

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0				0	@	P	`	p				ー	タ	ミ		
1			!	1	A	Q	a	q			。	ア	チ	ム		
2			"	2	B	R	b	r			「	イ	ツ	メ		
3			#	3	C	S	c	s			」	ウ	テ	モ		
4			\$	4	D	T	d	t			、	エ	ト	ヤ		
5			%	5	E	U	e	u			・	オ	ナ	ユ		
6			&	6	F	V	f	v			ラ	カ	ニ	ヨ		
7			'	7	G	W	g	w			ァ	キ	ヌ	ラ		
8			(	8	H	X	h	x			ィ	ク	ネ	リ		
9			)	9	I	Y	i	y			ゥ	ケ	ノ	ル		
A			*	:	J	Z	j	z			エ	コ	ハ	レ		
B			+	;	K	[	k	{			オ	サ	ヒ	ロ		
C			,	<	L	¥	l				ヤ	シ	フ	ワ		
D			-	=	M	]	m	}			ユ	ス	ヘ	ン		
E			.	>	N	^	n	~			ヨ	セ	ホ	・		
F			/	?	O	_	o				ッ	ソ	マ	*		

語と言語

$\mathcal{A}$: アルファベット (文字集合)

語 (word) $\mathcal{A}$ に属する文字を 0 個以上並べた文字列 (string)

$$w = w_1 w_2 \dots w_n (w_i \in \mathcal{A})$$

ただし、0 個からなる文字列も語: **空語** (empty word) ε

言語 (language) 語の集合 $\mathcal{L} = \{w\}$

ただし、適当に文字を並べても言語にはならない

apple $\in$ 英語、 aelpp $\notin$ 英語

$\mathcal{A}$ 上の全ての語全体 $\mathcal{A}^*$

$$\mathcal{A}^* = \bigcup_{i=0}^{\infty} \mathcal{A}^i = \mathcal{A}^0 + \mathcal{A}^1 + \mathcal{A}^2 + \mathcal{A}^3 + \dots, \quad \mathcal{A}^0 = \{\varepsilon\}$$

一般には言語 $\mathcal{L}$ は無限個の語 w の集まりからなる:

$$\#\{w \mid w \in \mathcal{L}\} = \infty$$

簡単な言語の例 (1) 正規言語

$$\mathcal{A} = \{0, 1\}$$

- $\mathcal{L}_1 = \{w \mid w \text{ は } 1 \text{ で始まり } 0 \text{ で終わる} \}$

$$\{10, 100, 110, 1000, 1010, 1100, 1110, \dots\}$$

- $\mathcal{L}_2 = \{w \mid w \text{ に含まれる } 1 \text{ の数が偶数} \}$

$$\{\varepsilon, 0, 00, \dots, 11, 011, 101, 110, \dots \\ 1111, 0111, 10111, 11011, 11101, 11110, \dots\}$$

- $\mathcal{L}_3 = \{w \mid w \text{ には } 1 \text{ の次の続けて } 1 \text{ が現れない} \}$

$$\{\varepsilon, 0, 1, 01, 10, 000, 001, 010, 100, 101, 0000, \dots\}$$

簡単な言語の例 (2) 文脈自由言語

$$\mathcal{A} = \{0, 1\}$$

- $\mathcal{L}_4 = \{0^n 1^n, | n \in \mathbb{N} = \{1, 2, 3, \dots\}\}$

$$\{01, 0011, 000111, 00001111, \dots\}$$

- $\mathcal{L}_5 = \{w, | w \text{ は } 0 \text{ と } 1 \text{ を同じ数だけ含む}\}$

$$\{\varepsilon, 01, 10, 0011, 0101, 0110, 1001, 1010, 1100, \dots\}$$

簡単な言語の例 (3) 非文脈自由言語

- $\mathcal{L}_6 = \{ww^R \mid w^R \text{ は } w \text{ を逆に書いた文字列}\}$

$$\{\varepsilon, 00, 11, 0000, 0110, 1001, 1111, \dots\}$$

- $\mathcal{L}_7 = \{a^n b^n c^n \mid n \in \mathbb{N} = \{1, 2, 3, \text{dots}\}\}$

$$\{abc, aabbcc, aaabbbccc, \dots\}$$

- $\mathcal{L}_8 = \{a^i b^j c^k \mid 0 \leq i \leq j \leq k\}$

- $\mathcal{L}_9 = \{ww \mid w \in \{0, 1\}^*\}$

言語の認識 (recognition)

与えられた文字列（語）がある言語に属しているかどうかを知ること

語がある言語に属しているか否かをどのように**判定**するのだろうか？

言語の生成 (generation)

ある言語に属している文字列をすべて作りだすこと

文法と**単語帳**（文の最小単位）さえあればできる？

言語の3つの定義と等価性

文字列（語）を**全て列挙**して言語を与える。
一般には、（可算）無限個の語を列挙しなければならない。

単語の全てと文を生成する**文法規則**全体。
自然言語では、その提示自体が莫大になる。人工言語なら比較的容易

ヒト（**機械**）が理解（**受理**）できる語の全体。

以上の3つは等価である：



自動販売機

- アルファベット $\mathcal{A} = \{10 \text{ 円硬貨}, 50 \text{ 円硬貨}, 100 \text{ 円硬貨}\}$
- 語 (文字列) = 投入するコイン列
- 受理 = 料金に達すること