形態素解析

参考：自然言語処理

長尾真

岩波講座ソフトウェア科学

データ型(1Byteで何が扱えるか)

- 1Byte=8bit

00000000
00000001
00000010
00000011
…
11111110
11111111

- 自然数として
  0～255
- 整数として
  −128～127
- 文字データ
  a,b,c,…,A,B,C,…,0,1,2,…
  ,.!”#$%&’()=-~^|¥`@{[}]+;:*
  <>?/_,…

文字データ(ASCII)

- *アルファベット大文字*

A,B,C,D,E,F,G,…,X,Y,Z

- *アルファベット小文字*

a,b,c,d,e,f,g,…,x,y,z

- *数字*

0,1,2,3,4,5,6,7,8,9

- *特殊文字*

!, “,#,$,%,&, ‘,(,),*,+,,,-,.,/,:,;,<,=,>,?,@,[,¥,],^,_, `,

{,|,},~

- *制御文字*

nul,soh,stx,etx,eot,enq,ack,bel,bs,ht,lf,vt,ff,cr,so,si,dle,dc1,

dc2,dc3,dc4,nak,syn,etb,can,em,sub,esc,fs,gs,rs,us,del

ASCII Code Table

例　Book

文字	二進数	16進数
‘B’	1000010	42
‘o’	1101111	6f
‘o’	1101111	6f
‘k’	1101011	6b

b6 b5 b4	000	001	010	011	100	101	110	111
b3–b0								
0000	*nul*	*dle*		0	@	P	`	p
0001	*soh*	*dc1*	!	1	A	Q	a	q
0010	*stx*	*dc2*	″	2	B	R	b	r
0011	*etx*	*dc3*	#	3	C	S	c	s
0100	*eot*	*dc4*	$	4	D	T	d	t
0101	*enq*	*nak*	%	5	E	U	e	u
0110	*ack*	*syn*	&	6	F	V	f	v
0111	*bel*	*etb*	’	7	G	W	g	w
1000	*bs*	*can*	(	8	H	X	h	x
1001	*ht*	*em*	)	9	I	Y	i	y
1010	*lf*	*sub*	*	:	J	Z	j	z
1011	*vt*	*esc*	+	;	K	[	k	{
1100	*ff*	*fs*	,	<	L	¥	l	|
1101	*cr*	*gs*	−	=	M	]	m	}
1110	*so*	*rs*	.	>	N	^	n	~
1111	*si*	*us*	/	?	O	_	o	*del*

例

- ‘a’=1100001(2)=97(10)=61(16)
- ‘b’=1100010(2) =98(10)=62(16)
- ‘1’=0110001(2)=49(10)=31(16)
- ‘ ’=0100000(2)=32(10)=20(16)
- ‘cr’=‘(改行)’=0001101(2)
  =13(10)=0D(16)

文書データの例

```
%cat txt
This is an example of text file.

%od -hc txt
0000000     5468    6973    2069    7320    616e    2065    7861    6d70
          T   h   i   s       i   s       a   n       e   x   a   m   p
0000020     6c65    206f    6620    7465    7874    2066    696c    652e
          l   e       o   f       t   e   x   t       f   i   l   e   .
0000040     0a00
          ¥n
0000041
```

日本語の扱い

- 漢字:最低数千文字
- 2Byte=16bitで1文字を指定する。
- 2^{16}=6.4万とおりの文字を扱える。
- いわゆる全角文字
- 1Byte文字（ASCII文字、半角文字）と混在。
- どう切り替えるか。

文字集合の例（JIS X 0208)

文字集合(JIS X 0208)

- 一つの文字を区(row.1~94)点(cell.1~94)で定義する。
  例:「医」＝16区69点（または16-69)
- 8,836(=94×94)字を扱うことが出来る。
- この構造は中国のGB2312、韓国のKS X 1001でも採用。

Row 16	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
00		亜	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪	握	渥	旭	葦	芦	鯵
20	梓	圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎	或	粟	袷	安	庵	按	暗	案	闇
40	鞍	杏	以	伊	位	依	偉	囲	夷	委	威	尉	惟	意	慰	易	椅	為	畏	異
60	移	維	緯	胃	萎	衣	謂	違	遺	医	井	亥	域	育	郁	磯	一	壱	溢	逸
80	稲	茨	芋	鰯	允	印	咽	員	因	姻	引	飲	淫	胤	蔭					

Row 17	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
00		院	陰	隠	韻	吋	右	宇	烏	羽	迂	雨	卯	鵜	窺	丑	碓	臼	渦	嘘
20	唄	鬱	蔚	鰻	姥	厩	浦	瓜	閏	噂	云	運	雲	荏	餌	叡	営	嬰	影	映
40	曳	栄	永	泳	洩	瑛	盈	穎	頴	英	衛	詠	鋭	液	疫	益	駅	悦	謁	越
60	閲	榎	厭	円	園	堰	奄	宴	延	怨	掩	援	沿	演	炎	焔	煙	燕	猿	縁
80	艶	苑	薗	遠	鉛	鴛	塩	於	汚	甥	凹	央	奥	往	応					

文字集合(JIS X 0208)

- 非漢字（524字）
  - 記号（1区、2区）：∴、∵、≦、¬、√、…
  - 英数字（3区）：A、B、C、…、1、2、…
  - ひらがな（4区）：あ、い、う、え、お、…
  - カタカナ（5区）：ア、イ、ウ、エ、オ、…
  - ギリシャ文字（6区）：α、β、γ、δ、…
  - キリル文字（7区）：А、Б、В、Г、Д、…
  - 罫線素片（8区）：┃、┏、┣、…
- 第一水準漢字（16区～47区。2965字）

  よく使われる漢字。五十音順（あいうえお順）に配置。
- 第二水準漢字（48区～83区。3384字。84区に6字）

  第一水準の漢字より、より使われない漢字。部首順に配置。

文字集合（Unicode）に対する文字符号化（UTF-8）

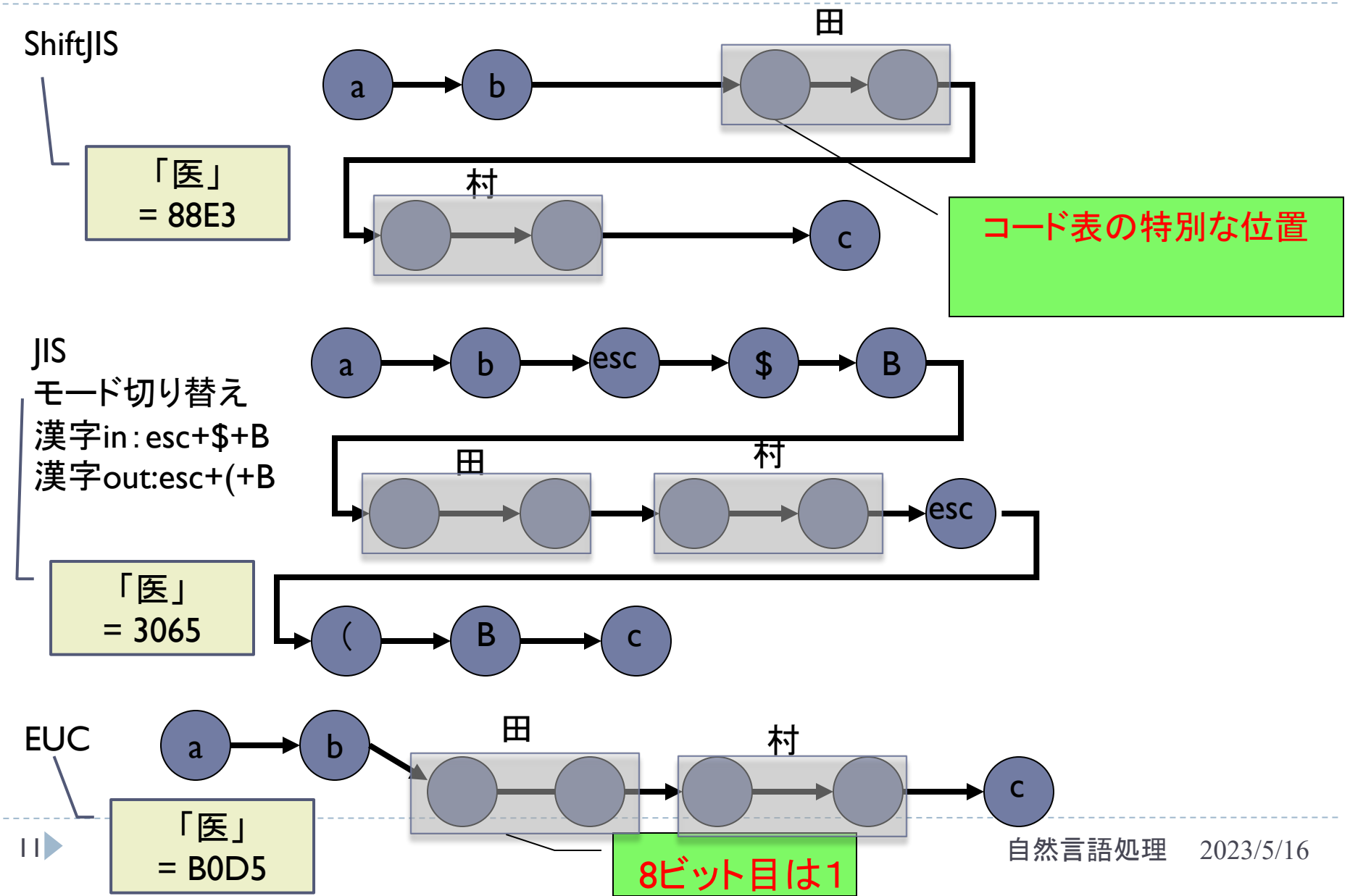
コンピュータで扱える文字の集合(Unicode)とその符号化(UTF-8)を
分離して考える。

文字集合	文字符号化
U+0000~U+007F	00~7F (7bit)
(ASCIIと同一）	
U+0080~U+07FF	C280~DFBF (11bit)
(非漢字の一部）	
U+0800~U+FFFF	E08080~EFBFBF (16bit)
(残りの漢字）	
U+10000~U+1FFFFF	F0808080~F7BFBFBF (21bit)
(第3、第4水準漢字）	

「医」
= U+533B

「医」
= E5 8C BB

コード化（EUC, SJIS, JIS）
「ab田村c」の各コード



形態素とは

自然言語の階層

- 音素

  意味の伝達における音の最小単位。

- 形態素

  意味を持つ最小の言語単位。一つ以上の音素から成る。

- 語

  文法上一つの機能を持つ最小の言語単位。一つ以上の形態素から成る。

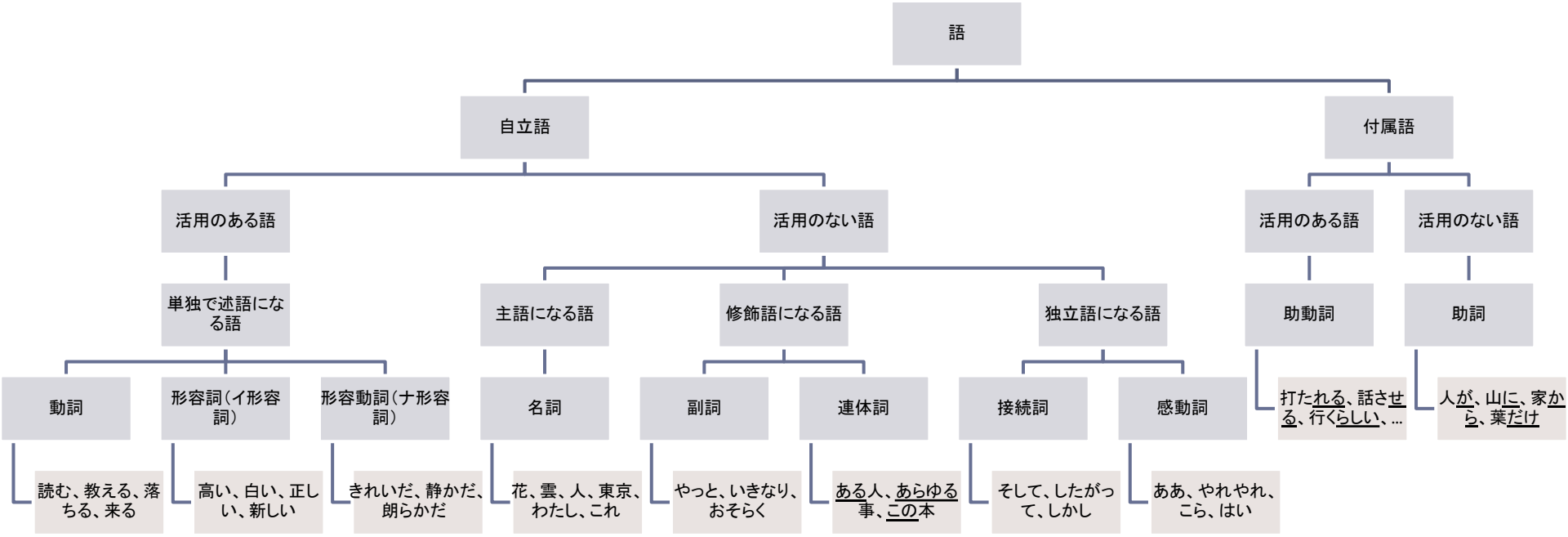
- 文

  ある内容を持ち形の上で完結した言語単位。一つ以上の語から成る。

- 文章・テキスト

  あるまとまった内容を表現するための文の順序づけられた集まり。一つ以上の文から成る。

日本語の品詞分類


語
自立語
付属語
活用のある語
活用のない語
活用のある語
活用のない語
単独で述語になる語
主語になる語
修飾語になる語
独立語になる語
助動詞
助詞
動詞
形容詞（イ形容詞）
形容動詞（ナ形容詞）
名詞
副詞
連体詞
接続詞
感動詞
打たれる、話させる、行くらしい、…
人が、山に、家から、葉だけ
読む、教える、落ちる、来る
高い、白い、正しい、新しい
きれいだ、静かだ、朗らかだ
花、雲、人、東京、わたし、これ
やっと、いきなり、おそらく
ある人、あらゆる事、この本
そして、したがって、しかし
ああ、やれやれ、こら、はい

形態素解析結果の例

```
いま　（いま）　いま　副詞
米国　（べいこく）　米国　地名
から　（から）　から　格助詞
は　（は）　は　副助詞
さまざまな　（さまざまな）　さまざまだ　形容詞　ナノ形容詞　ダ列基本連体
警報　（けいほう）　警報　普通名詞
が　（が）　が　格助詞
発せ　（はっせ）　発する　動詞　サ変動詞　文語未然形
られて　（られて）　られる　動詞性接尾辞　母音動詞　タ系連用テ形
いる　（いる）　いる　動詞性接尾辞　母音動詞　基本形
。　（。）　。　句点
EOS
商務　（しょうむ）　商務　普通名詞
省　（しょう）　省　普通名詞
の　（の）　の　接続助詞
日本　（にっぽん）　日本　地名
など　（など）　など　副助詞
鉄鋼　（てっこう）　鉄鋼　普通名詞
メーカー　（めーかー）　メーカー　普通名詞
製品　（せいひん）　製品　普通名詞
に　（に）　に　格助詞
対する　（たいする）　対する　動詞　サ変動詞　基本形
ダンピング　（ダンピング）　ダンピング　カタカナ
仮　（かり）　仮　普通名詞
決定　（けってい）　決定　サ変名詞
や　（や）　や　接続助詞
上院　（じょういん）　上院　普通名詞
……
EOS
```

形態素解析用の単語辞書と連接可能性辞書

基本形：読(よ)む
語幹：読
未然形：ま、も　eg.読まない
連用形：み、ん　eg.読みながら
終止形：む　eg.読む
連体形：む　eg.読むとき
仮定形：め　eg.読めば、
命令形：め　eg.読め！

単語辞書

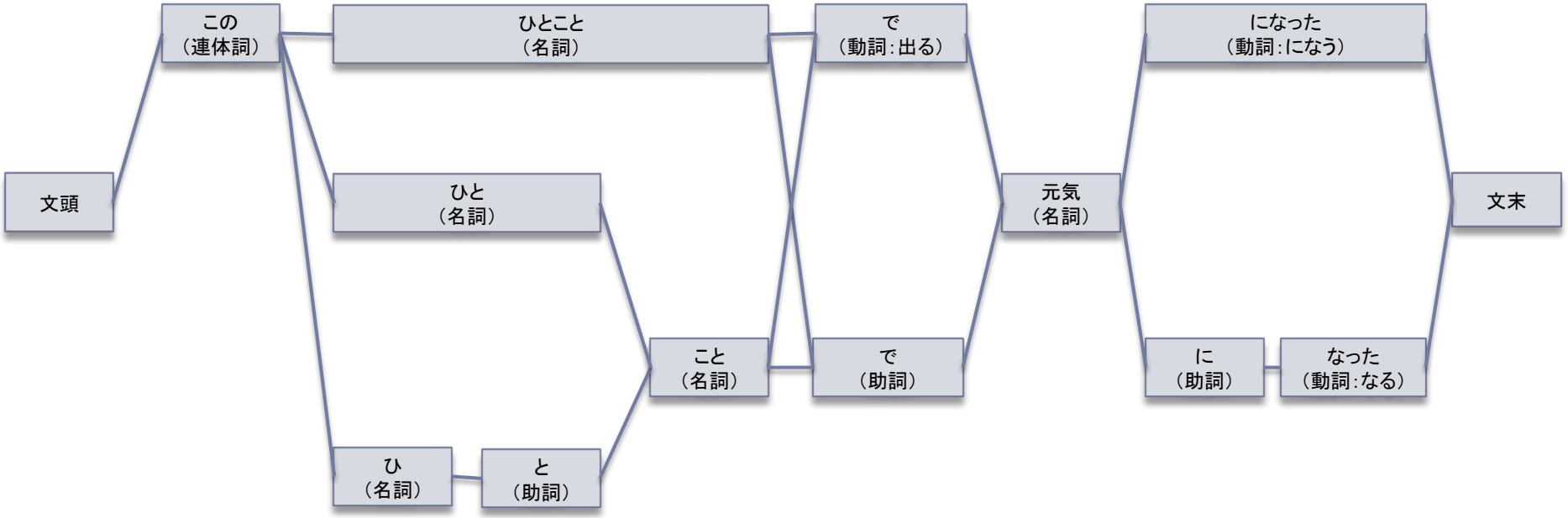
見出し語	読み	品詞	活用形
日本語	にほんご	名詞：普通名詞	-
英語	えいご	名詞：普通名詞	-
読む	よむ	動詞	子音動詞マ行
書く	かく	動詞	子音動詞カ行
…	…	…	…

接続可能性辞書

左側	右側
文頭	名詞、動詞、形容詞、形容動詞、副詞、連体詞、接続詞、感動詞
名詞 動詞　連用形	助詞：格助詞、副助詞、引用助詞、名詞接続助詞 助詞：終助詞　さ、か、かしら、ね、な、よ、なあ
動詞　基本形	助詞：終助詞　ぞ、ぜ、わ
動詞　意志形	助詞：終助詞　ぜ
動詞　基本形	助詞：述語接続助詞　し、が、とも、から、けれども、なら
動詞　連用形	助詞：述語接続助詞　つつ、ながら
動詞　未然形	接尾辞：動詞性接尾辞　れる、られる
・・・	
基本形、命令形、名詞、助詞：終助詞	文末

2単語間の連接規則による形態素解析結果（制約）

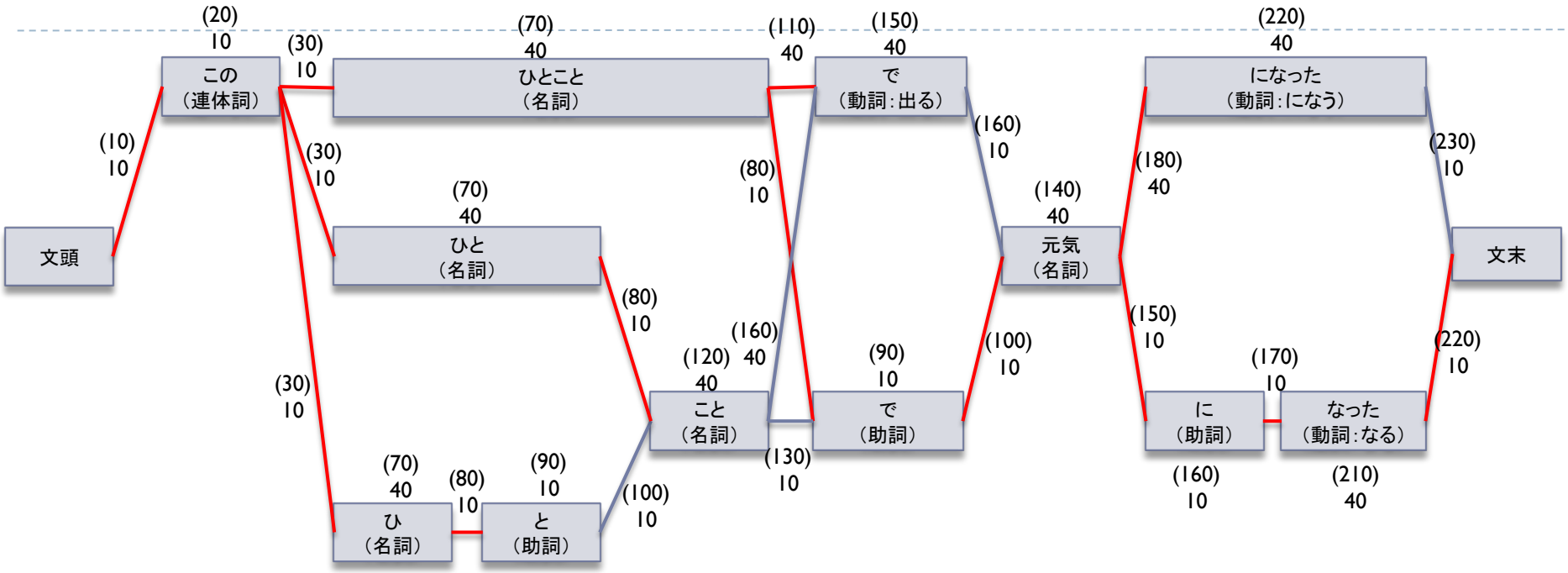
例文:「このひとことで元気になった」



形態素解析における優先規則

- 最長一致法
  長い形態素を優先
- 2文節最長一致法
  2文節ごとの長さが長い解を優先
- 形態素数最小法
  形態素の数が少ない方を優先
- 文節数最小法
  文節数の少ない解を優先
- コスト最小法
  語や語の連接にコストを与え、総コストの少ない解を優先

Viterbiアルゴリズムによるコスト計算（優先規則あるいは選好）



(累計)
品詞コストor接続コスト

Viterbiアルゴリズム

1. 品詞コスト、接続コストを求める
2. 「文頭」からのコストを累計する
   複数の入力経路があるときは、一番コストが低い経路を選ぶ
3. 「文末」から「文頭」に向かって経路を選択する

英語の形態素解析

- 入力の単語列

$$w_1, w_2, \cdots, w_n$$

- 求めたい品詞列

$$C_1, C_2, \cdots, C_n \quad \text{s.t. } \mathrm{P}(C_1, C_2, \cdots, C_n | w_1, w_2, \cdots, w_n) \text{が最大}$$

- ベイズの定理により変形

$$\frac{\mathrm{P}(C_1, C_2, \cdots, C_n) \times \mathrm{P}(w_1, w_2, \cdots, w_n | C_1, C_2, \cdots, C_n)}{\mathrm{P}(w_1, w_2, \cdots, w_n)}$$

- 分母は品詞と無関係なので無視
- 分子第1項：直前の品詞のみに依存すると仮定

$$\mathrm{P}(C_1, C_2, \cdots, C_n) \cong \prod_{i=1}^{n} \mathrm{P}(C_i | C_{i-1})$$

ただしC_0は文頭マーク

- 分子第2項：前後の品詞とは独立と仮定

$$\mathrm{P}(w_1, w_2, \cdots, w_n | C_1, C_2, \cdots, C_n) \cong \prod_{i=1}^{n} \mathrm{P}(w_i | C_i)$$

- 目的関数

$$\prod_{i=1}^{n} \mathrm{P}(C_i | C_{i-1}) \times \mathrm{P}(w_i | C_i)$$

bigram確率

カテゴリC_iにおけるw_iの確率

サンプルコーパス中の単語/品詞の出現確率

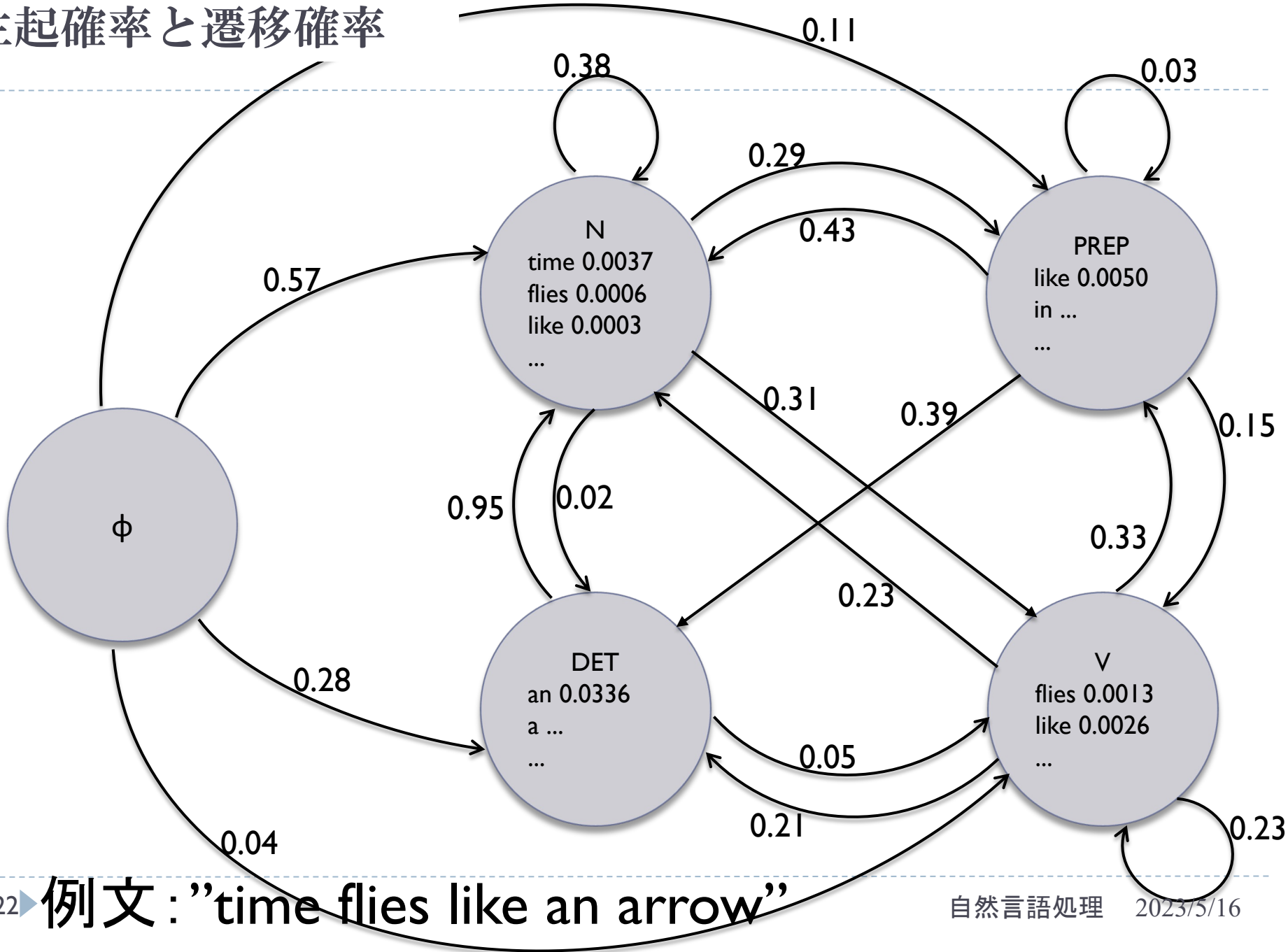
wの出現頻度とP(w|C)

w \ C	N		V		DET		PREP	
time	13	0.0037	0	0.0000	0	0.0000	0	0.0000
flies	2	0.0006	2	0.0013	0	0.0000	0	0.0000
like	1	0.0003	4	0.0026	0	0.0000	7	0.005
an	0	0.0000	0	0.0000	37	0.0336	0	0.0000
arrow	1	0.0003	0	0.0000	0	0.0000	0	0.0000
…	…		…		…		…	
Total	3481	1.0000	1515	1.0000	1102	1.0000	1405	1.0000

サンプルコーパス中の品詞のbigram

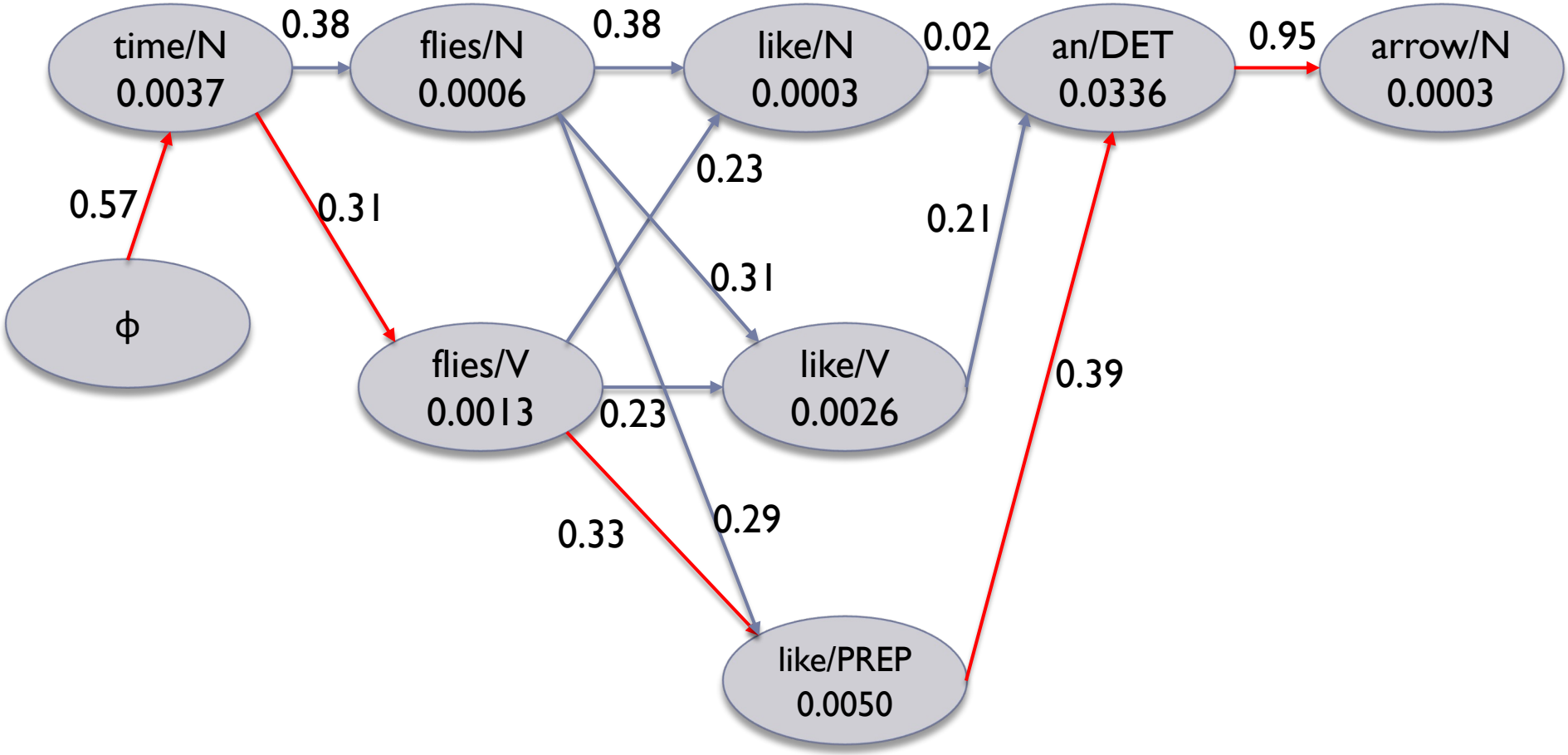
C_i \ C_{i-1}	ϕ		N		V		DET		PREP	
N	392	0.57	1111	0.38	326	0.23	1050	0.95	605	0.43
V	28	0.04	918	0.31	313	0.23	52	0.05	204	0.15
DET	194	0.28	78	0.02	289	0.21	0	0.00	541	0.39
PREP	71	0.11	840	0.29	456	0.33	0	0.00	38	0.03
Total	685	1.00	2947	1.00	1384	1.00	1102	1.00	1388	1.00

隠れマルコフ・モデル
生起確率と遷移確率



例文：”time flies like an arrow”

Viterbiアルゴリズムによる品詞付け



例文：”time flies like an arrow”

中国語の
形態素解析システム（その１）

1. **Jieba（结巴中文分词）**

   https://github.com/fxsjy/jieba

   よく使われているオープンソースの形態素解析ツール、

   java c++ pythonなど多数のプログラミング言語で利用可

   能である。カスタム辞書、繁体字もサポートしている。

2. **SnowNLP**

   https://github.com/isnowfy/snownlp

   Pythonで書かれた中国語向けの形態素解析システム。単

   語の感情分析、テキスト分類、TextRankによる文章要約

   の機能も実装されている。

中国語の
形態素解析システム（その２）

3. HanLP

https://www.hanlp.com/

パワフルな中国語形態素解析システム、ロカールかRESTful API(有料)で利用可能。Word2vec、固有表現抽出、データ・クラスタリング、文章要約など自然言語処理用の機能がたくさん実装されている。

4. 哈工大ltp

https://ltp.ai/

ハルビン工業大学(哈尔滨工业大学)が開発したオープンソースの形態素解析システム。PYTHON，JAVA，Rust, C++での利用は可能。RESTful APも提供されている。

オンラインデモ：http://ltp.ai/demo.html

ほかにも百度NLP、盘古、Yaha、 清华THULACなど多数あり

まとめ

1. 日本語の扱い(コード)
2. 日本語の品詞
3. 形態素解析の原理
    1. Viterbiアルゴリズム
4. 英語の形態素解析
5. 中国語の形態素解析

課題

1. 「このひとことで元気になった」
2. 「time flies like an arrow」

上記1，2のそれぞれについて、形態素解析の全部の解を求めよ。

bi-gram確率（接続コスト）と生起確立（品詞コスト）を単純に加算してよい。