

การสร้างคลังข้อมูลภาษาแบบ DIY

- Electronic Texts การ represent ภาษาในคอมพิวเตอร์
- Format แบบต่างๆ กับการเก็บข้อมูลภาษา
- การ copy texts จากแหล่งต่างๆ
- การใช้โปรแกรม OCR & Scanner
- การแปลงข้อมูลจากไฟล์ format ต่างๆ
- การใช้โปรแกรม Htrack copy website
- การใช้โปรแกรม DeHtml แปลงเป็น plain text

Electronic Texts

- E-text = ข้อมูลภาษาที่เก็บอยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้
- ข้อดี : ใช้คอมพิวเตอร์เพื่อเข้าถึงข้อมูลได้รวดเร็ว แม่นยำ
- เก็บข้อมูลภาษาเป็น E-text ได้อย่างไร
- ถ้าพูดถึงเฉพาะภาษาเขียนหรือตัวอักษรที่ถูกเก็บในคอมพิวเตอร์ คอมพิวเตอร์เป็นอุปกรณ์อิเล็กทรอนิกส์ สิ่งที่มีมันรู้จักคือ on หรือ off คือหน่วยที่เล็กที่สุดในคอมพิวเตอร์ที่เราเรียกว่า bit ค่าของ bit มีได้เพียงแค่ 0 หรือ 1
- เปรียบเหมือนกับในโลกของคอมพิวเตอร์มี phoneme เพียง 2 ตัว การที่จะ represent หน่วยที่ใหญ่ขึ้นก็ต่อองอาศัย phoneme มากกว่า 1 หน่วย ในทางคอมพิวเตอร์ เขาใช้ 8 bit รวมเป็นหน่วยที่ใหญ่ขึ้นเรียกว่า byte ซึ่งจะสามารถใช้ represent สิ่งที่แตกต่างกันได้ถึง $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 256$ แบบ หรือเรียกเป็น code จาก 0-255 ได้

E-texts

- Character encoding เริ่มตั้งแต่การใช้ morse code ใช้รหัส on, off และความยาวของการเคาะ แทนตัวอักษร เช่น _ . แทน A, _ . . . แทน B, _ . _ . แทน C เป็นต้น
- ต่อมามีการใช้ code แบบ 5-bit (Baudot code) 6-bit (Hollerith code)
- ต่อมา มี 7-bit ASCII code ในปี 1967 (ได้ 128 อักษร),
ISO ยอมรับเป็นมาตรฐานในปี 1972
ASCII = American Standard Code for Information Interchange
มีตัวอักษรที่พิมพ์ได้ 95 ตัวเริ่มจากรหัส 32 ถึง 126.
- ต่อมา มีการ extend ASCII เป็น 8-bit เพื่อให้มี code แทนอักษรภาษาต่างประเทศอื่น เป็นกลุ่ม ISO 8859
- ISO 8859-1 Latin1 ใช้สำหรับภาษายุโรปตะวันตก เช่น French, Italian, German, Spanish ISO 8859-11 ใช้สำหรับภาษาไทย

ASCII

Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	\$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(	56	8	72	H	88	X	104	h	120	x
41	)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[	107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	
45	-	61	=	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	-
48	0	64	@	80	P	96	`	112	p		

ASCII

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	í	¢	£	¤	¥	¦	§		©	ª	«	¬	­	®	¯
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF
ø	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

- ตัวอักษรที่เพิ่มเติมมาใน ISO-8859-1 หรือ Latin-1



ISO-8859-x	Name	Year	Languages covered
1	Latin-1	1987	Western European languages
2	Latin-2	1987	East European languages
3	Latin-3	1988	Southern European languages
4	Latin-4	1988	Northern European languages
5	Latin/Cyrillic	1988	Russian, Bulgarian, Ukrainian, etc.
6	Latin/Arabic	1987	Arabic
7	Latin/Greek	1987	Greek
8	Latin/Hebrew	1988	Hebrew
9	Latin-5	1989	Turkish (Replaces Latin-3)
10	Latin-6		Northern European languages (Unifies Latin-1 and Latin-4)
11	Latin/Thai		Thai
12	Currently unassigned		May be used in future for Indian or Vietnamese
13	Latin-7		Baltic languages (Replaces Latin-4 and supplements Latin-6)
14	Latin-8		Celtic characters
15	Latin-9	1998	Western European languages (Replaces Latin-1 and adds the euro symbol plus a few missing French and Finnish characters)
16	Latin-10		Eastern European languages (Replaces Latin-2 and adds the euro symbol plus a few missing Romanian characters)

				b ₈	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
				b ₇	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	
				b ₆	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	
				b ₅	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	
				00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15		
b ₄	b ₃	b ₂	b ₁	00			SP	0	๑	P	`	p			NBSP	ฐ	ภ	ะ	เ	๐	0
			01				!	1	A	Q	a	q			ก	ท	ม	็	แ	๑	1
			02				"	2	B	R	b	r			ข	ฌ	ย	า	โ	๒	2
			03				#	3	C	S	c	s			ช	ฌ	ร	ำ	ใ	๓	3
			04				\$	4	D	T	d	t			ค	ค	ฤ	็	ไ	๔	4
			05				%	5	E	U	e	u			ค	ค	ล	็	า	๕	5
			06				&	6	F	V	f	v			ฌ	ถ	ภ	็	า	๖	6
			07				'	7	G	W	g	w			ง	ท	ว	็	็	๗	7
			08				(	8	H	X	h	x			จ	ธ	ศ	ุ	็	๘	8
			09				)	9	I	Y	i	y			ฉ	น	ษ	ู	็	๙	9
			10				*	:	J	Z	j	z			ช	บ	ส	ุ	็	๑	A
			11				+	;	K	[	k	{			ช	ป	ห		็	๒	B
			12				,	<	L	\	l				ฌ	ผ	พ		็	๓	C
			13				-	=	M	]	m	}			ญ	ฝ	อ		็	๔	D
			14				.	>	N	^	n	~			ญ	พ	ฮ		็	๕	E
			15				/	?	O	_	o				ญ	ฟ	๑	฿	๖	๖	F
				0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	hex	

E-texts

- เดิมรหัสภาษาไทยไม่มีมาตรฐาน มี รหัส KU, RW ต่อมา มี TIS-620
- ISO 8859-11 Thai encoding พัฒนามาจาก TIS-620 (สมอ) (เพิ่ม A0 สำหรับ non-breaking space)
- Microsoft เองก็มี encoding สำหรับภาษาต่างๆ ภาษาไทย Windows-874
- หลายภาษา เช่น จีน ญี่ปุ่น 8-bit code นั้นไม่พอ มีการใช้ 16-bit และมี encoding แบบต่างๆ เช่น JIS C 6220, JIS X 0213, CNS 11463, ... ($2^{16} = 65,536$ แบบ)

E-texts

- มีหลายกลุ่มพยายามสร้าง encoding ให้ใช้ได้กับทุกภาษา
- มาตรฐานการ encoding ที่สำคัญปัจจุบัน คือ unicode. (range U+0000..U+10FFFF)
- implementation ของ unicode ทำออกมาหลาย character encoding เช่น UTF-8 (1-4 8 bit bytes), UTF-16 (1-2 16 bit code unit²), UTF-32 (32 bit code unit)
- สามารถแปลงข้อมูลจาก UTF หนึ่งไปอีกแบบได้ (UTF = unicode transformation format)

E-texts

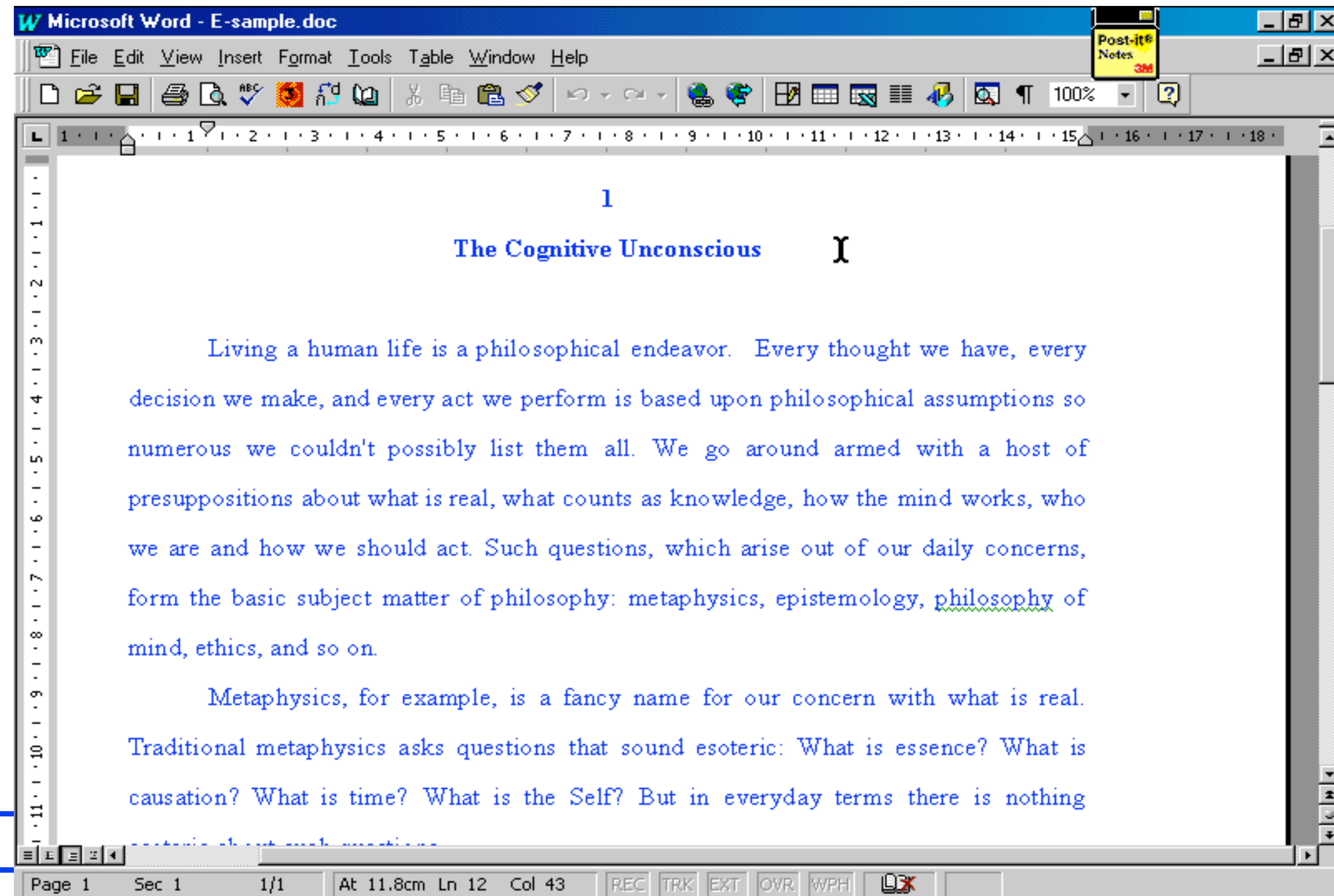
- UTF-8 (8-bit UCS/Unicode Transformation Format) เป็น variable-length character encoding (1-4 bytes), backwards compatible กับ ASCII.
- The first 128 characters (US-ASCII) need one byte.
- The next 1920 characters need two bytes to encode. This includes Latin letters with diacritics and characters from Greek, Cyrillic, Coptic, Armenian, Hebrew, Arabic, Syriac and Tāna alphabets.
- Three bytes are needed for the rest of the Basic Multilingual Plane (which contains virtually all characters in common use).
- Four bytes are needed for characters in the other planes of Unicode, which include less common CJK characters and various historic scripts
<http://unicode-search.net/unicode-namesearch.pl?term=thai>

	0E0	0E1	0E2	0E3	0E4	0E5	0E6	0E7
0		จ ๐๐10	ภ ๐๐20	๐๐ ๐๐30	ไ ๐๐40	๐ ๐๐50		
1	ก ๐๐11	ท ๐๐11	ม ๐๐21	๐๐ ๐๐31	แ ๐๐41	๑ ๐๐51		
2	ข ๐๐12	ฅ ๐๐12	ย ๐๐22	า ๐๐32	โ ๐๐42	๒ ๐๐52		
3	ช ๐๐13	ณ ๐๐13	ร ๐๐23	ำ ๐๐33	ใ ๐๐43	๓ ๐๐53		
4	ค ๐๐14	ด ๐๐14	ฤ ๐๐24	๐๐ ๐๐34	ไ ๐๐44	๔ ๐๐54		
5	ก ๐๐15	ต ๐๐15	ล ๐๐25	๐๐ ๐๐35	า ๐๐45	๕ ๐๐55		
6	ฆ ๐๐16	ถ ๐๐16	ภ ๐๐26	๐๐ ๐๐36	ง ๐๐46	๖ ๐๐56		
7	ง ๐๐17	ท ๐๐17	ว ๐๐27	๐๐ ๐๐37	๐๐ ๐๐47	๗ ๐๐57		
8	จ ๐๐18	ธ ๐๐18	ศ ๐๐28	๐๐ ๐๐38	๐๐ ๐๐48	๘ ๐๐58		
9	ฉ ๐๐19	น ๐๐19	ษ ๐๐29	๐๐ ๐๐39	๐๐ ๐๐49	๙ ๐๐59		
A	ช ๐๐1A	บ ๐๐1A	ส ๐๐2A	๐๐ ๐๐3A	๐๐ ๐๐4A	๑๐ ๐๐5A		
B	ช ๐๐1B	ป ๐๐1B	ท ๐๐2B		๐๐ ๐๐4B	๑๑ ๐๐5B		
C	ฅ ๐๐1C	ผ ๐๐1C	ฬ ๐๐2C		๐๐ ๐๐4C			
D	ญ ๐๐1D	ฝ ๐๐1D	อ ๐๐2D		๐๐ ๐๐4D			
E	ฎ ๐๐1E	พ ๐๐1E	ฮ ๐๐2E		๐๐ ๐๐4E			
F	ฏ ๐๐1F	ฟ ๐๐1F	า ๐๐2F	฿ ๐๐3F	๐๐ ๐๐4F			

ประเภทของ E-texts

- Plain text เป็นแฟ้มข้อมูลที่เก็บเฉพาะ code ที่แทนตัวอักษรในภาษาเท่านั้น จะไม่มี code พิเศษอื่นๆที่จะบอก format เช่น การใช้ Notepad
- Formatted text เป็นแฟ้มข้อมูลที่เก็บ information อื่นนอกเหนือจากตัว text ด้วย เช่น เก็บ format หรือรหัสพิเศษที่จะกำหนดรูปร่างหน้าตา text เมื่อจะพิมพ์ออกมา text ชนิดนี้ได้จากการใช้โปรแกรม word processor ทั่วไป เช่น MS Word, AmiPro, Word Perfect. etc. นอกจากนี้แล้ว file บาง file อาจอยู่ในรูปแบบเฉพาะเพื่อการพิมพ์หรือแสดงหน้าจอเท่านั้น เช่น Tex-Latex files (*.TEX), Postscript files (*.PS), Acrobat file (*.PDF), etc.)
- Annotated texts จะเหมือนกับ formatted text ในแง่ที่ตัว text มี information อื่นนอกเหนือจากตัวภาษาอยู่ใน file นั้นด้วย แต่ information ใน annotated text ไม่ได้มีเพื่อการจัดรูปแบบการพิมพ์ แต่เป็น information ที่ describe ตัว text นั้นเอง เช่น LOB corpus, CSAE, etc.

Examples of E-texts



Plain Text

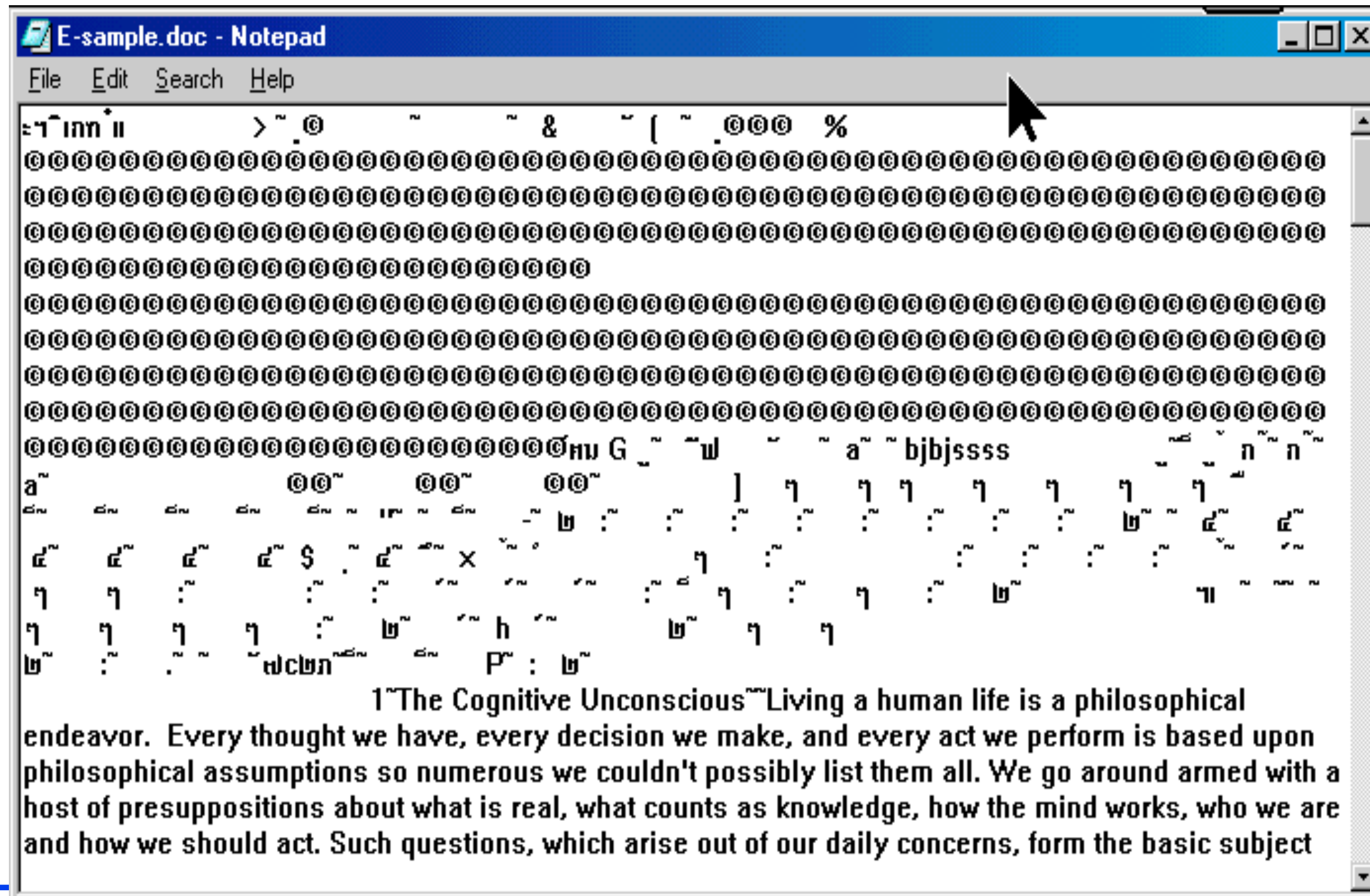
- 1
- The Cognitive Unconscious
- Living a human life is a philosophical endeavor. Every thought we have, every decision we make, and every act we perform is based upon philosophical assumptions so numerous we couldn't possibly list them all. We go around armed with a host of presuppositions about what is real, what counts as knowledge, how the mind works, who we are and how we should act. Such questions, which arise out of our daily concerns, form the basic subject matter of philosophy: metaphysics, epistemology, philosophy of mind, ethics, and so on.
- Metaphysics, for example, is a fancy name for our concern with what is real. Traditional metaphysics asks questions that sound esoteric: What is essence? What is causation? What is time? What is the Self? But in everyday terms there is nothing esoteric about such questions.

Hex

00000000: 31 0d 0a 54 68 65 20 43 6f 67 6e 69 74 69 76 65
00000010: 20 55 6e 63 6f 6e 73 63 69 6f 75 73 0d 0a 0d 0a
00000020: 4c 69 76 69 6e 67 20 61 20 68 75 6d 61 6e 20 6c
00000030: 69 66 65 20 69 73 20 61 20 70 68 69 6c 6f 73 6f
00000040: 70 68 69 63 61 6c 20 65 6e 64 65 61 76 6f 72 2e
00000050: 20 20 45 76 65 72 79 20 74 68 6f 75 67 68 74 20
00000060: 77 65 20 68 61 76 65 2c 20 65 76 65 72 79 20 64
00000070: 65 63 69 73 69 6f 6e 20 77 65 20 6d 61 6b 65 2c
00000080: 20 61 6e 64 20 65 76 65 72 79 20 61 63 74 20 77
00000090: 65 20 70 65 72 66 6f 72 6d 20 69 73 20 62 61 73
000000a0: 65 64 20 75 70 6f 6e 20 70 68 69 6c 6f 73 6f 70
000000b0: 68 69 63 61 6c 20 61 73 73 75 6d 70 74 69 6f 6e
000000c0: 73 20 73 6f 20 6e 75 6d 65 72 6f 75 73 20 77 65
000000d0: 20 63 6f 75 6c 64 6e 27 74 20 70 6f 73 73 69 62
000000e0: 6c 79 20 6c 69 73 74 20 74 68 65 6d 20 61 6c 6c
000000f0: 2e 20 57 65 20 67 6f 20 61 72 6f 75 6e 64 20 61
00000100: 72 6d 65 64 20 77 69 74 68 20 61 20 68 6f 73 74
00000110: 20 6f 66 20 70 72 65 73 75 70 70 6f 73 69 74 69
00000120: 6f 6e 73 20 61 62 6f 75 74 20 77 68 61 74 20 69
00000130: 73 20 72 65 61 6c 2c 20 77 68 61 74 20 63 6f 75
00000140: 6e 74 73 20 61 73 20 6b 6e 6f 77 6c 65 64 67 65
00000150: 2c 20 68 6f 77 20 74 68 65 20 6d 69 6e 64 20 77
00000160: 6f 72 6b 73 2c 20 77 68 6f 20 77 65 20 61 72 65
00000170: 20 61 6e 64 20 68 6f 77 20 77 65 20 73 68 6f 75
00000180: 6c 64 20 61 63 74 2e 20 53 75 63 68 20 71 75 65
00000190: 73 74 69 6f 6e 73 2c 20 77 68 69 63 68 20 61 72
000001a0: 69 73 65 20 6f 75 74 20 6f 66 20 6f 75 72 20 64
000001b0: 61 69 6c 79 20 63 6f 6e 63 65 72 6e 73 2c 20 66
000001c0: 6f 72 6d 20 74 68 65 20 62 61 73 69 63 20 73 75
000001d0: 62 6a 65 63 74 20 6d 61 74 74 65 72 20 6f 66 20
000001e0: 70 68 69 6c 6f 73 6f 70 68 79 3a 20 6d 65 74 61
000001f0: 70 68 79 73 69 63 73 2c 20 65 70 69 73 74 65 6d
00000200: 6f 6c 6f 67 79 2c 20 70 68 69 6c 6f 73 6f 70 68
00000210: 79 20 6f 66 20 6d 69 6e 64 2c 20 65 74 68 69 63
00000220: 73 2c 20 61 6e 64 20 73 6f 20 6f 6e 2e 20 0d 0a

1..The Cognitive
Unconscious....
Living a human l
ife is a philoso
phical endeavor.
Every thought
we have, every d
ecision we make,
and every act w
e perform is bas
ed upon philosop
hical assumption
s so numerous we
couldn't possib
ly list them all
. We go around a
rmed with a host
of presuppositi
ons about what i
s real, what cou
nts as knowledge
, how the mind w
orks, who we are
and how we shou
ld act. Such que
stions, which ar
ise out of our d
aily concerns, f
orm the basic su
bject matter of
philosophy: meta
physics, epistem
ology, philosoph
y of mind, ethic
s, and so on. ..

MS Word



Hex

00000470:	00 00 3a 01 00 00 00 00 00 00 00 f2 01 00 00 00 00	...:.....ò.....
00000480:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000490:	00 00 fa 00 00 00 0e 00 00 00 00 08 01 00 00 0e 00	..ú.....
000004a0:	00 00 e6 00 00 00 00 00 00 00 00 e6 00 00 00 00 00	..æ.....æ.....
000004b0:	00 00 e6 00 00 00 00 00 00 00 00 e6 00 00 00 00 00	..æ.....æ.....
000004c0:	00 00 3a 01 00 00 00 00 00 00 00 f2 01 00 00 00 00	...:.....ò.....
000004d0:	00 00 8a 01 00 00 68 00 00 00 8a 01 00 00 00 00	..Š...h...Š.....
000004e0:	00 00 00 00 00 00 00 00 00 00 00 f2 01 00 00 00 00	ò.....
000004f0:	00 00 e6 00 00 00 00 00 00 00 00 e6 00 00 00 00 00	..æ.....æ.....
00000500:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000510:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000520:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000530:	00 00 00 00 00 00 00 00 00 00 00 f2 01 00 00 00 00	ò.....
00000540:	00 00 3a 01 00 00 00 00 00 00 2e 01 00 00 0c 00	...:.....
00000550:	00 00 00 20 98 f7 63 f2 c0 01 16 01 00 00 00 00	... ÷còÀ.....
00000560:	00 00 16 01 00 00 00 00 00 00 50 01 00 00 3a 00	P...:.
00000570:	00 00 f2 01 00 00 00 00 00 00 00 00 00 00 00 00	..ò.....
00000580:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000590:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005a0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005b0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005c0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005d0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005e0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
000005f0:	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	
00000600:	31 0d 54 68 65 20 43 6f 67 6e 69 74 69 76 65 20	1.The Cognitive
00000610:	55 6e 63 6f 6e 73 63 69 6f 75 73 0d 0d 4c 69 76	Unconscious..Liv
00000620:	69 6e 67 20 61 20 68 75 6d 61 6e 20 6c 69 66 65	ing a human life
00000630:	20 69 73 20 61 20 70 68 69 6c 6f 73 6f 70 68 69	is a philosophi
00000640:	63 61 6c 20 65 6e 64 65 61 76 6f 72 2e 20 20 45	cal endeavor. E
00000650:	76 65 72 79 20 74 68 6f 75 67 68 74 20 77 65 20	very thought we
00000660:	68 61 76 65 2c 20 65 76 65 72 79 20 64 65 63 69	have, every deci
00000670:	73 69 6f 6e 20 77 65 20 6d 61 6b 65 2c 20 61 6e	sion we make, an
00000680:	64 20 65 76 65 72 79 20 61 63 74 20 77 65 20 70	d every act we p

HTML

- `<HTML>`
- `<HEAD>`
- `<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=windows-1252">`
- `<TITLE>1</TITLE>`
- `</HEAD>`
- `<BODY>`
- `<B><FONT COLOR="#0000ff"><P ALIGN="CENTER">1</P>`
- `<P ALIGN="CENTER">The Cognitive Unconscious</P>`
- `</B></FONT>`
- `<P ALIGN="JUSTIFY">Living a human life is a philosophical endeavor. Every thought we have, every decision we make, and every act we perform is based upon philosophical assumptions so numerous we couldn't possibly list them all. We go around armed with a host of presuppositions about what is real, what counts as knowledge, how the mind works, who we are and how we should act. Such questions, which arise out of our daily concerns, form the basic subject matter of philosophy: metaphysics, epistemology, philosophy of mind, ethics, and so on. </P>`

Hex

00000000: 3c 48 54 4d 4c 3e 0d 0a 3c 48 45 41 44 3e 0d 0a
00000010: 3c 4d 45 54 41 20 48 54 54 50 2d 45 51 55 49 56
00000020: 3d 22 43 6f 6e 74 65 6e 74 2d 54 79 70 65 22 20
00000030: 43 4f 4e 54 45 4e 54 3d 22 74 65 78 74 2f 68 74
00000040: 6d 6c 3b 20 63 68 61 72 73 65 74 3d 77 69 6e 64
00000050: 6f 77 73 2d 31 32 35 32 22 3e 0d 0a 3c 54 49 54
00000060: 4c 45 3e 31 3c 2f 54 49 54 4c 45 3e 0d 0a 3c 2f
00000070: 48 45 41 44 3e 0d 0a 3c 42 4f 44 59 3e 0d 0a 0d
00000080: 0a 3c 42 3e 3c 46 4f 4e 54 20 43 4f 4c 4f 52 3d
00000090: 22 23 30 30 30 30 66 66 22 3e 3c 50 20 41 4c 49
000000a0: 47 4e 3d 22 43 45 4e 54 45 52 22 3e 31 3c 2f 50
000000b0: 3e 0d 0a 3c 50 20 41 4c 49 47 4e 3d 22 43 45 4e
000000c0: 54 45 52 22 3e 54 68 65 20 43 6f 67 6e 69 74 69
000000d0: 76 65 20 55 6e 63 6f 6e 73 63 69 6f 75 73 3c 2f
000000e0: 50 3e 0d 0a 3c 2f 42 3e 3c 2f 46 4f 4e 54 3e 0d
000000f0: 0a 3c 46 4f 4e 54 20 43 4f 4c 4f 52 3d 22 23 30
00000100: 30 30 30 66 66 22 3e 3c 50 20 41 4c 49 47 4e 3d
00000110: 22 4a 55 53 54 49 46 59 22 3e 4c 69 76 69 6e 67
00000120: 20 61 20 68 75 6d 61 6e 20 6c 69 66 65 20 69 73
00000130: 20 61 20 70 68 69 6c 6f 73 6f 70 68 69 63 61 6c
00000140: 20 65 6e 64 65 61 76 6f 72 2e 20 20 45 76 65 72
00000150: 79 20 74 68 6f 75 67 68 74 20 77 65 20 68 61 76
00000160: 65 2c 20 65 76 65 72 79 20 64 65 63 69 73 69 6f
00000170: 6e 20 77 65 20 6d 61 6b 65 2c 20 61 6e 64 20 65
00000180: 76 65 72 79 20 61 63 74 20 77 65 20 70 65 72 66
00000190: 6f 72 6d 20 69 73 20 62 61 73 65 64 20 75 70 6f
000001a0: 6e 20 70 68 69 6c 6f 73 6f 70 68 69 63 61 6c 20
000001b0: 61 73 73 75 6d 70 74 69 6f 6e 73 20 73 6f 20 6e
000001c0: 75 6d 65 72 6f 75 73 20 77 65 20 63 6f 75 6c 64
000001d0: 6e 27 74 20 70 6f 73 73 69 62 6c 79 20 6c 69 73
000001e0: 74 20 74 68 65 6d 20 61 6c 6c 2e 20 57 65 20 67
000001f0: 6f 20 61 72 6f 75 6e 64 20 61 72 6d 65 64 20 77
00000200: 69 74 68 20 61 20 68 6f 73 74 20 6f 66 20 70 72
00000210: 65 73 75 70 70 6f 73 69 74 69 6f 6e 73 20 61 62
00000220: 6f 75 74 20 77 68 61 74 20 69 73 20 72 65 61 6c

<HTML>..<HEAD>..
<META HTTP-EQUIV
="Content-Type"
CONTENT="text/ht
ml; charset=wind
ows-1252">..<TIT
LE>1</TITLE>..</
HEAD>..<BODY>...
.<FONT COLOR=
"#0000ff"><P ALI
GN="CENTER">1</P
>..<P ALIGN="CEN
TER">The Cogniti
ve Unconscious</
P>....
.<FONT COLOR="#0
000ff"><P ALIGN=
"JUSTIFY">Living
a human life is
a philosophical
endeavor. Ever
y thought we hav
e, every decisio
n we make, and e
very act we perf
orm is based upo
n philosophical
assumptions so n
umerous we could
n't possibly lis
t them all. We g
o around armed w
ith a host of pr
esuppositions ab
out what is real

PDF

00000000:	25	50	44	46	2d	31	2e	32	20	0d	0a	25	e2	e3	cf	d3
00000010:	0d	0a	20	0d	0a	31	30	20	30	20	6f	62	6a	0d	0a	3c
00000020:	3c	0d	0a	2f	4c	65	6e	67	74	68	20	31	31	20	30	20
00000030:	52	0d	0a	2f	46	69	6c	74	65	72	20	2f	46	6c	61	74
00000040:	65	44	65	63	6f	64	65	20	0d	0a	3e	3e	0d	0a	73	74
00000050:	72	65	61	6d	0d	0a	48	89	8d	56	c9	92	db	38	0c	fd
00000060:	82	fe	07	dc	66	a9	6e	85	da	a5	5c	ba	b2	56	4d	d5
00000070:	4c	e5	10	4f	e5	4c	cb	b4	c5	b4	44	6a	48	c9	1e	ff
00000080:	fd	00	5c	d4	72	e2	c3	1c	ba	4d	49	24	1e	80	f7	00
00000090:	f0	fd	ee	21	67	39	54	6d	0b	b0	fb	f8	c0	80	41	0a
000000a0:	e6	04	0f	6f	3e	e3	2a	c3	97	47	7c	09	bb	0e	e8	ff
000000b0:	05	7e	4d	7f	83	dd	f7	87	a7	aa	4a	ea	12	9e	b2	82
000000c0:	8e	c1	13	4b	58	d9	14	7e	5f	5c	e2	e6	5d	2f	e0	83
000000d0:	3e	29	39	cb	b3	80	bf	55	a7	95	ed	a4	5e	ac	37	92
000000e0:	a6	2c	c9	d0	4a	d1	38	2b	6f	3e	a7	11	d2	19	2c	bd
000000f0:	c1	34	29	fd	12	0d	fe	29	cf	52	9d	80	43	bf	8c	5c
00000100:	c1	20	8f	02	a4	c5	e7	a9	97	83	b6	1a	7f	3a	3e	80
00000110:	50	07	c1	cf	da	24	00	9f	ce	c2	5c	61	ee	f5	72	ea
00000120:	67	b8	08	e8	f9	59	3c	82	a0	d7	de	8b	bc	ba	89	a3
00000130:	62	31	8e	26	65	01	f6	20	3a	69	a5	56	74	7e	e4	2f
00000140:	78	9e	ab	83	b7	01	bc	73	66	27	61	8e	da	8c	e4	cd
00000150:	9e	5b	71	80	65	c2	fd	b7	6e	71	6b	97	71	9a	d1	90
00000160:	05	ab	1d	3a	bb	c1	2e	f2	da	61	e7	49	dd	ba	25	62
00000170:	ab	65	14	06	53	46	20	9d	5e	86	83	fa	65	86	49	5b
00000180:	2b	f7	c3	15	33	60	67	8c	4e	8c	c0	87	21	81	6f	02
00000190:	4e	1a	38	ee	47	ff	b8	19	d1	8f	8b	9c	7b	4a	98	c6
000001a0:	8d	fa	e8	40	77	bf	3b	b4	26	ad	62	a4	4d	9b	7b	b4
000001b0:	c9	08	bb	4c	68	5e	7a	37	f9	5e	2f	18	5f	cf	67	8a
000001c0:	cc	08	3e	3c	fa	27	74	45	cd	f8	dd	c2	8b	d2	97	41
000001d0:	1c	4e	98	95	5e	5f	c8	19	18	25	c2	5f	b4	79	b1	b4
000001e0:	5b	6f	41	b3	ba	09	ac	56	7e	89	a0	18	19	37	c2	e5
000001f0:	94	2c	e0	a3	ed	29	52	ca	6d	02	5f	97	ae	87	7f	16
00000200:	61	9d	47	64	50	e2	0b	6e	a4	15	40	ce	e9	23	fe	18

%PDF-1.2 ..%âãĬŎ
.. ..10 0 obj..
<../Length 11 0
R../Filter /Flat
eDecode ..>>..st
ream..H%ŲVÉ' Ū8.ý
,p.Üf@n...ÚŴ\ °²VMŌ
Lă.OăLE' Ă'DjHE.ý
ý.\ÔrăĂ.°MI\$.€÷.
šýî!g9Tm.°ûøÀ€A.
æ..o>ă*Ă-G|.»..èÿ
.~M[]fÝ÷÷\$²Jê.ž²,
ŽÁ.KXŮ.~_ \ âæ]/àf
>)9Ě³€ç UŠ·íx^¬7'
|,ÉĐJÑ8+o>\$.Ō.,½
Á4) ý..p) ĨR]€CçE\
Á [] .xĂç©—f¶.[]:>€
P.ÁİÚ\$.ŸÎÂ\ aiöré
g..èùY<, xB<¼°%£
b1Ž&e.ö :iŴVt~ä/
xŽ«f..¼sf'aŽÚČăÍ
ž[q€eĂý.nqk—qšŃ
.«.:»Ă.òÚaçIÝ°%b
«e..SF []^†fú†I[
÷÷Ă.3`gENCEÀ# ![]o.
N.8îGŷ,.Ň[]<œ{J~Æ
[]úè@wç;´ &-bxM>{´
É.»Lh^z7ù^/. _İgŠ
ì.><ú'tEÍøÝĂ<Ō-A
.N~·^_È..%Ă_´y±´
[oA³°.¬V~¾ ..7Ăă
~,à£í) RÊm._—@#[].
a[]GdPĂ.nx. @İé#p.
8p%ù@...uĂà¶-fQg.3

การสร้าง E-Texts

- พิมพ์เอง
- Scan และใช้โปรแกรม OCR (optical character recognition)
 - โปรแกรมการ Scan มักจะ integrate เข้ากับโปรแกรมต่างๆ เช่น Photoshop, Omnipage ได้โดยตรง
 - สิ่งที่ได้คือ image เป็นข้อมูลภาพของจุดสีต่างๆ
 - แปลงเป็น text ด้วยโปรแกรม OCR แล้วเก็บผลที่ได้เป็น text หรือ format อื่นๆ
 - แปลง pdf เป็น text ด้วยการ save เป็น plain text ถ้าไม่ได้ นำเข้า Google Drive แปลงเป็น doc

การสร้าง E-Text

- ใช้ file ที่มีผู้สร้างไว้แล้ว
 - Oxford Text Archive, Project Gutenberg
 - Text จาก Online Book Pages,
 - download จาก internet
 - save ทีละ page, ทีละ file, ใช้ Save-as
 - website copy : โปรแกรม HTTrack

การ Extract text จากเว็บ

- <http://boilerpipe-web.appspot.com/> เป็น web api
- <http://corpus.tools/wiki/Justext> Python library & Online demo
- <http://www.octoparse.com/> โปรแกรมบน Windows
- <http://parsehub.com/>
- ใช้ Chrome Extension เช่น Advanced Web Scraper เลือกข้อมูลที่ต้องการ ตั้งชื่อ field บันทึกข้อมูลเป็น csv, json

พระราชบัญญัติลิขสิทธิ์พ.ศ. ๒๕๓๗

- "ทำซ้ำ" หมายความว่า ความรวมถึง คัดลอกไม่ว่าด้วยวิธีใด ๆ เลียนแบบ ทำสำเนา ทำ แม่พิมพ์ บันทึกเสียง บันทึกภาพ หรือบันทึกเสียงและภาพ จากต้นฉบับ จากสำเนาหรือจากการโฆษณา ในส่วนอันเป็นสาระสำคัญ ทั้งนี้ ไม่ว่าทั้งหมดหรือบางส่วน สำหรับในส่วนที่เกี่ยวกับ โปรแกรมคอมพิวเตอร์ให้หมายความถึงคัดลอกหรือทำสำเนาโปรแกรมคอมพิวเตอร์จากสื่อบันทึกใด ไม่ว่าด้วยวิธีใด ๆ ในส่วน อันเป็นสาระสำคัญ โดยไม่มีลักษณะเป็นการจัดทำงานขึ้นใหม่ ทั้งนี้ ไม่ว่าทั้งหมดหรือบางส่วน
- "ดัดแปลง" หมายความว่า ทำซ้ำโดยเปลี่ยนรูปใหม่ ปรับปรุงแก้ไข เพิ่มเติม หรือ จำลองงานต้นฉบับ ในส่วนอันเป็นสาระสำคัญโดยไม่มีลักษณะเป็นการจัดทำงานขึ้นใหม่ ทั้งนี้ ไม่ว่า ทั้งหมดหรือบางส่วน

- "เผยแพร่ต่อสาธารณชน" หมายความว่า ทำให้ปรากฏต่อสาธารณชน โดย การแสดง การบรรยาย การสวด การบรรเลงการทำให้ปรากฏด้วยเสียงและหรือภาพ การก่อสร้าง การจำหน่าย หรือโดยวิธีอื่นใดซึ่งงานที่ได้จัดทำขึ้น
- มาตรา ๗ สิ่งต่อไปนี้ไม่ถือว่าเป็นงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้
 - (๑)ข่าวประจำวัน และข้อเท็จจริงต่าง ๆ ที่มีลักษณะเป็นเพียงข่าวสารอันมิใช่งานในแผนกวรรณคดี แผนกวิทยาศาสตร์ หรือแผนกศิลปะ
 - (๒)รัฐธรรมนูญ และกฎหมาย
 - (๓)ระเบียบ ข้อบังคับ ประกาศ คำสั่ง คำชี้แจง และหนังสือตอบโต้ของ กระทรวง ทบวง กรม หรือหน่วยงานอื่นใดของรัฐหรือของท้องถิ่น
 - (๔)คำพิพากษา คำสั่ง คำวินิจฉัย และรายงานของทางราชการ
 - (๕)คำแปลและการรวบรวมสิ่งต่าง ๆ ตาม (๑) ถึง (๔) ที่กระทรวง ทบวง กรม หรือหน่วยงานอื่นใดของรัฐหรือของท้องถิ่นจัดทำขึ้น

- มาตรา ๑๙ ภายใต้บังคับมาตรา ๒๑ และมาตรา ๒๒ ลิขสิทธิ์ตามพระราชบัญญัตินี้ให้มียุ่ตลอดอายุของผู้สร้างสรรค์และมียุ่ต่อไปอีกเป็นเวลา ห้าสิบปี นับแต่ผู้สร้างสรรค์ถึงแก่ความตายในกรณีที่มีผู้ส่ร้างสรรค์ร่วมลิขสิทธิ์ในงานดังกล่าวให้มียุ่ตลอดอายุของผู้ส่ร้างสรรค์ร่วม และมียุ่ต่อไปอีกเป็นเวลาห้าสิบปีนับแต่ผู้ส่ร้างสรรค์ร่วมคนสุดท้ายถึงแก่ความตาย
ถ้าผู้ส่ร้างสรรค์หรือผู้ส่ร้างสรรค์ร่วมทุกคนถึงแก่ความตายก่อนที่จะได้มีการ โฆษณางานนั้น ให้ลิขสิทธิ์ดังกล่าวมีอายุห้าสิบปีนับแต่ได้มีการ โฆษณาเป็นครั้งแรก
- มาตรา ๒๗ การกระทำอย่างใดอย่างหนึ่งแก่งานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ โดยไม่ ได้รับอนุญาตตามมาตรา ๑๕ (๕) ให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้กระทำดังต่อไปนี้
 - (๑) ทำซ้ำหรือดัดแปลง
 - (๒) เผยแพร่ต่อสาธารณชน

- มาตรา ๓๒ การกระทำแก่งานอันมีลิขสิทธิ์ของบุคคลอื่นตามพระราชบัญญัตินี้ หากไม่ขัดต่อการแสวงหาประโยชน์จากงานอันมีลิขสิทธิ์ตามปกติของเจ้าของลิขสิทธิ์ และไม่กระทบกระเทือนถึงสิทธิอันชอบด้วยกฎหมายของเจ้าของลิขสิทธิ์เกินสมควร มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ภายใต้บังคับบทบัญญัติในวรรคหนึ่ง การกระทำอย่างหนึ่งอย่างใดแก่งานอันมีลิขสิทธิ์ตามวรรคหนึ่ง มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้กระทำได้ดังต่อไปนี้
 - (๑) วิจัยหรือศึกษางานนั้น อันมิใช่การกระทำเพื่อหากำไร
 - (๒) ใช้เพื่อประโยชน์ของตนเอง หรือเพื่อประโยชน์ของตนเองและบุคคลอื่น ในครอบครัวหรือญาติสนิท
 - (๓) ดิจม วิจารณ์ หรือแนะนำผลงาน โดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น
 - (๔) เสนอรายงานข่าวทางสื่อสารมวลชน โดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น
 - (๕) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ เพื่อประโยชน์ในการพิจารณาของศาลหรือเจ้าพนักงานซึ่งมีอำนาจตามกฎหมาย หรือในการรายงานผลการพิจารณาดังกล่าว

- (๖) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ โดยผู้สอนเพื่อประโยชน์ ในการสอนของตน อันมิใช่การกระทำให้ปรากฏ
- (๗) ทำซ้ำ ดัดแปลงบางส่วนของงาน หรือตัดทอน หรือทำบทสรุปโดยผู้สอน หรือสถาบันศึกษาเพื่อแจกจ่ายหรือจำหน่ายแก่ผู้เรียน ในชั้นเรียนหรือในสถาบันศึกษา ทั้งนี้ ต้องไม่เป็นการกระทำเพื่อหากำไร
- (๘) นำงานนั้นมาใช้เป็นส่วนหนึ่งในการถามและตอบในการสอบ
- มาตรา ๓๓ การกล่าว คัดลอก เลียน หรืออ้างอิงงานบางตอนตามสมควรจากงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ โดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้ปฏิบัติตามมาตรา ๓๒ วรรคหนึ่ง
- มาตรา ๓๔ การทำซ้ำโดยบรรณารักษ์ของห้องสมุดซึ่งงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์หากการทำซ้ำนั้นมิได้มีวัตถุประสงค์เพื่อหากำไร และได้ปฏิบัติตามมาตรา ๓๒ วรรคหนึ่ง ในกรณีดังต่อไปนี้
 - (๑) การทำซ้ำเพื่อใช้ในห้องสมุดหรือให้แก่ห้องสมุดอื่น
 - (๒) การทำซ้ำงานบางตอนตามสมควรให้แก่บุคคลอื่น เพื่อประโยชน์ในการ วิจัยหรือการศึกษา

พรบ.ลิขสิทธิ์ (ฉบับที่ 2) 2558

- เพิ่มเติมเรื่อง
 - “ข้อมูลการบริหารสิทธิ”
 - ข้อมูลที่บ่งชี้ถึงผู้สร้างสรรค์ งานสร้างสรรค์ นักแสดง การแสดง เจ้าของลิขสิทธิ์ หรือระยะเวลาและเงื่อนไขการใช้งานอันมีลิขสิทธิ์ ตลอดจนตัวเลขหรือรหัสแทนข้อมูลดังกล่าว โดยข้อมูลเช่นนี้ติดอยู่หรือปรากฏเกี่ยวข้องกับงานอันมีลิขสิทธิ์หรือสิ่งบันทึกการแสดง
 - “มาตรการทางเทคโนโลยี”
 - เทคโนโลยีที่ออกแบบมาเพื่อป้องกันการทำซ้ำหรือควบคุมการเข้าถึงงานอันมีลิขสิทธิ์หรือสิ่งบันทึกการแสดง โดยเทคโนโลยีเช่นนี้ได้นำมาใช้กับงาน อันมีลิขสิทธิ์หรือสิ่งบันทึกการแสดงนั้นอย่างมีประสิทธิภาพ
 - “การหลบเลี่ยงมาตรการทางเทคโนโลยี”
 - การกระทำด้วยประการใด ๆ ที่ทำให้มาตรการทางเทคโนโลยีไม่เกิดผล

Types of Corpora

- Written corpora vs spoken corpora
- General reference corpora vs Special purpose corpora
- Sample corpora vs full text corpora
- Static corpora vs monitor corpora
- Monolingual corpora vs multilingual corpora
- Parallel corpora vs comparable corpora
- Learner corpora vs native corpora
- Plain corpora vs annotated corpora

BNC Sample - annotated corpus

▷ [BNC](#) ▷ [docs](#) ▷ [eg.xml](#)



What do BNC texts look like 'in the raw'? Like other XML files they are understandable, provided that you know what the tags are for and how to interpret the syntax. The markup used in the BNC is documented in the User Reference Guide, which also gives usage examples for all the markup constructs contained you will encounter in the real corpus. However, for the insatiably curious, we present here a complete BNC file, the smallest in the corpus: text FX8, containing nine "utterances" from a doctor's surgery, with a total of 128 words...

```
<bncDoc xml:id="FX8"><teiHeader><fileDesc><titleStmt><title> General practitioner's surgery: medical consultation. Sample<
<stext type="OTHERSP"><u who="FX8PSUNK">
<s n="1"><w c5="ITJ" hw="ah" pos="INTERJ">Ah </w><w c5="AV0" hw="there" pos="ADV">there </w><w c5="PNP" hw="we" pos="PRON"
<s n="2"><w c5="AV0" hw="right" pos="ADV">Right </w><unclear/><w c5="AJ0" hw="abdominal" pos="ADJ">abdominal </w><w c5="I
<s n="3"><w c5="PNP" hw="she" pos="PRON">She </w><w c5="VDD" hw="do" pos="VERB">did</w><w c5="XX0" hw="not" pos="ADV">n't
<s n="4"><w c5="UNC" hw="erm" pos="UNC">Erm </w><w c5="PNP" hw="she" pos="PRON">she </w><w c5="VBD" hw="be" pos="VERB">was
<s n="5"><align with="FX8LC001"/><w c5="CJC" hw="and" pos="CONJ">And </w><w c5="PNP" hw="you" pos="PRON">you </w><unclear/>
<s n="6"><w c5="PNP" hw="she" pos="PRON">She </w><w c5="VVD" hw="say" pos="VERB">said </w><w c5="PNP" hw="she" pos="PRON">
<s n="7"><w c5="PNP" hw="she" pos="PRON">She</w><w c5="VBZ" hw="be" pos="VERB">'s </w><unclear/><w c5="AV0" hw="here" pos=
<s n="8"><pause/><w c5="AV0" hw="so" pos="ADV">So </w><w c5="PNP" hw="I" pos="PRON">I </w><w c5="VDB" hw="do" pos="VERB">
<s n="9"><unclear/><w c5="PNP" hw="it" pos="PRON">it</w><w c5="VBZ" hw="be" pos="VERB">'s </w><w c5="AV0" hw="just" pos="I
<s n="10"><unclear/><c c5="PUN">?</c></s></u><u who="PS22T">
<s n="11"><w c5="ITJ" hw="yeah" pos="INTERJ">Yeah</w><c c5="PUN">.</c></s></u><u who="FX8PS000">
<s n="12"><w c5="AV0" hw="okay" pos="ADV">Okay</w><c c5="PUN">.</c></s></u><u who="PS22T">
<s n="13"><w c5="ITJ" hw="yeah" pos="INTERJ">Yeah</w><c c5="PUN">.</c></s></u><u who="FX8PS000">
<s n="14"><w c5="UNC" hw="erm" pos="UNC">erm</w><c c5="PUN">,</c><w c5="ORD" hw="first" pos="ADJ">first </w><unclear/><w
<s n="15"><w c5="UNC" hw="erm" pos="UNC">Erm </w><unclear/><w c5="DT0" hw="this" pos="ADJ">this </w><w c5="PNI" hw="one"
```