LTI
Language, Technology and the Internet

Writing as Language Technology

Francis Bond

Division of Linguistics and Multilingual Studies

http://www3.ntu.edu.sg/home/fcbond/

bond@ieee.org

Lecture 2

Overview

➢ The origins of writing

➢ Different writing systems

➢ Representing writing on computers

➢ Writing versus talking

The Origins of Writing

➤ Writing was invented independently in at least three places:

  ➤ Mesopotamia
  ➤ China
  ➤ Mesoamerica

  Probably also Egypt (Earliest Egyptian Glyphs) and possibly the Indus valley.

➤ The written records are incomplete

➤ Gradual development from pictures/tallies

Follow the money

➤ Before 2700, writing is only accounting.

    ➤ Temple and palace accounts
    ➤ Gold, Wheat, Sheep

➤ How it developed in Mesopotamia

    ➤ One token per thing (in a clay envelope)
    ➤ One token per thing in the envelope and marked on the outside
    ➤ One mark per thing
    ➤ One mark and a symbol for the number
    ➤ Finally symbols for names

Denise Schmandt-Besserat (1997) *How writing came about*. University of Texas Press

Clay Tokens and Envelope
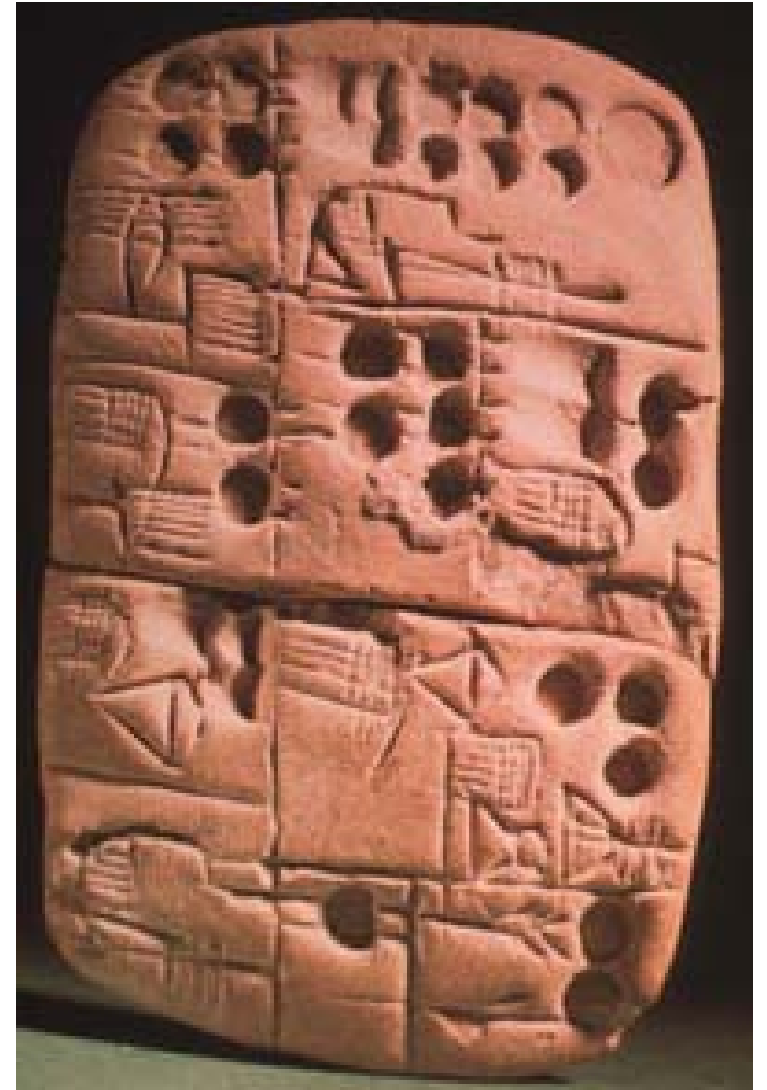
Clay Tablet

Writing systems used for human languages

➤ What is writing?

> A system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer. *Peter T. Daniels, The World's Writing Systems*

➤ Different types of writing systems are used:

   ➤ Alphabetic
   ➤ Syllabic
   ➤ Logographic

What is represented?

In English:

➤ Phonemes: /maɪ dɒg laɪks ˈɒrɪndʒɪz/ (45)

➤ Syllables: maɪ dɒg laɪks (ˈɒr)(ɪn)(dʒɪz) (10,000+)

➤ Morphemes: my/me+'s dog like+s orange+s (100,000+)

➤ Words: *my dog likes oranges* (200,000+)

➤ Concepts: `speaker poss dog`$_{canine}$`:SG fond orange`$_{fruit}$`:PL` (400,000++)

1. Some languages have fewer syllables

2. It is hard to visually distinguish 1,000+ different symbols, …

Alphabetic systems

➤ Alphabets (phonemic alphabets)

    ➤ represent all sounds, i.e., consonants and vowels

    ➤ Examples: Etruscan, Latin, Cyrillic, Runic, International Phonetic Alphabet, ?Korean

➤ Abjads (consonant alphabets)

    ➤ represent consonants only (sometimes plus selected vowels; vowel diacritics generally available)

    ➤ Examples: Arabic, Aramaic, Hebrew
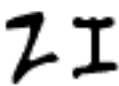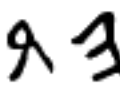
Alphabet example: Russian
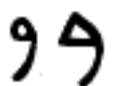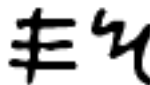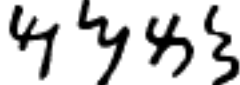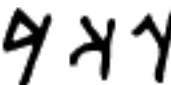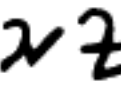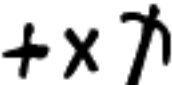
The Cyrillic alphabet is used to write many languages, mainly Slavic. Here is the set used for Russian.

А а	Б б	В в	Г г	Д д	Е е	Ё ё	Ж ж	З з	И и	Й й
а	бэ	вэ	гэ	дэ	е	ё	жэ	зэ	и	и краткое
a	b	v	g	d	e	ë	ž	z	i	j
[a]	[b]	[v]	[g]	[d]	[je/ie/e/ɛ]	[jo/ʲo/o]	[ʒ]	[z]	[i]	[j]

К к	Л л	М м	Н н	О о	П п	Р р	С с	Т т	У у	Ф ф
ка	эль	эм	эн	о	пэ	эр	эс	тэ	у	эф
k	l	m	n	o	p	r	s	t	u	f
[k]	[l]	[m]	[n]	[o]	[p]	[r]	[s]	[t]	[u]	[f]

Х х	Ц ц	Ч ч	Ш ш	Щ щ	Ъ ъ	Ы ы	ь	Э э	Ю ю	Я я
ха	цэ	че	ша	ща	твёрдый знак	ы	мягкий знак	э	ю	я
kh/h/x	c	č	š	šč	″	y	′	ė	ju	ja
[x]	[ts]	[tʃ]	[ʃ]	[ʃtʃ/ʃʃ]	–	[ɨ]	[ʲ]	[ɛ]	[ju/ʲu]	[ja/ʲa]

(from: http://www.omniglot.com/writing/russian.htm)

Abjad example: Phoenician

An abjad used to write Phoenician, created between the 18th and 17th centuries BC; assumed to be the forerunner of the Greek and Hebrew alphabet.

ḥēt	zayin	wāw	hē	dālet	gīmel	bēt	ʾālef
ḥ	z	w	h	d	g	b	ʾ
sāmek	nun	mēm		lāmed	kaf	yōd	ṭēt
s	n	m		l	k	y	ṭ
tāw	śin/šin		rēš	qōf	ṣādē	pē	ʿayin
t	š		r	q	ṣ	p	ʿ

(from: http://www.omniglot.com/writing/phoenician.htm)

A note on the letter-sound correspondence

➢ Alphabets use letters to encode sounds (consonants, vowels).

➢ But the correspondence between spelling and pronounciation in many languages is quite complex, i.e., not a simple one-to-one correspondence.

➢ Example: English

  ➢ same spelling – different sounds: ough: *ought, cough, tough, through, though, hiccough*
  ➢ silent letters: *knee, knight, knife, debt, psychology, mortgage*
  ➢ one letter – multiple sounds: *exit*
  ➢ multiple letters – one sound: *the, revolution*
  ➢ alternate spellings: *jail* or *gaol*

Capitalization Can Carry Meaning

- ➤ *Die Spinnen!* "The spiders!"
  - ➤ *Die spinnen!* "They are crazy!"

- ➤ *Er hatte liebe Genossen.* "He had kind companions."
  - ➤ *Er hatte Liebe genossen.* "He had enjoyed love."

- ➤ *Sich brüsten und Anderem zuwenden.* "to gloat and turn towards other things"
  - ➤ *Sich Brüsten und Anderem zuwenden.* "to turn towards breasts and other things"

- ➤ *Der Gefangene floh.* "The prisoner escaped."
  - ➤ *Der gefangene Floh.* "The imprisoned flea"

Syllabic systems

➢ Syllabic alphabets (Alphasyllabaries)

    ➢ writing systems with symbols that represent a consonant with a vowel, but the vowel can be changed by adding a diacritic (= a symbol added to the letter).

    ➢ Examples: Balinese, Javanese, Tibetan, Tamil, Thai, Tagalog

➢ Syllabaries

    ➢ writing systems with separate symbols for each syllable of a language

    ➢ Examples: Cherokee. Ethiopic, Cypriot, Ojibwe, Hiragana (Japanese)

Syllabic alphabet example: Lao

Script developed in the 14th century to write the Lao language, based on an early version of the Thai script, which was developed from the Old Khmer script, which was itself based on Mon scripts.

Example for vowel diacritics around the letter k:

ກະ	ກິ	ກຸ	ກຶ	ກາ	ກີ	ກູ	ກື	ເກະ	ແກະ
ka	ki	ku	ku'	ka:	ki:	ku:	ku:'	ke	kae
[ka]	[ki]	[ku]	[kɯ]	[ka:]	[ki:]	[ku:]	[kɯ:]	[ke]	[kae]

ໂກະ	ເກ	ແກ	ໂກ	ເກາະ	ເກິ	ເກຶ	ເກຍ	ກົວ	ເກີ
ko	ke:	kae:	ko:	ko'	koe	kia	kia	kua	koe:y
[ko]	[ke:]	[kæ]	[ko:]	[kɔ]	[kɤ]	[kiə]	[kiə]	[kuə]	[kɤ:j]

ເກີຍ	ກໍ	ເກິ	ເກືອ	ເກົາ	ໃກ	ໄກ	ກໍາ	ກ໌
koe:y	ko':	koe:	ku'a	kaw	kay	kay	kam	k
[kɤ:j]	[kɔ:]	[kɤ:]	[kuə]	[kaw]	[kaj]	[kaj]	[kam]	[k]

Syllabic alphabet example: Hiragana

Script developed in 10th century from Chinese characters. 52 characters.

平仮名 （ひらがな） hiragana

a	あ 安	i	い 以	u	う 宇	e	え 衣	o	お 於
ka	か 加	ki	き 幾	ku	く 久	ke	け 計	ko	こ 己
sa	さ 左	shi	し 之	su	す 寸	se	せ 世	so	そ 曽
ta	た 太	chi	ち 知	tsu	つ 川	te	て 天	to	と 止
na	な 奈	ni	に 仁	nu	ぬ 奴	ne	ね 祢	no	の 乃
ha	は 波	hi	ひ 比	fu	ふ 不	he	へ 部	ho	ほ 保
ma	ま 末	mi	み 美	mu	む 武	me	め 女	mo	も 毛
ya	や 也			yu	ゆ 由			yo	よ 与
ra	ら 良	ri	り 利	ru	る 留	re	れ 礼	ro	ろ 呂
wa	わ 和	wi	ゐ 為			we	ゑ 恵	wo	を 遠
								n	ん 无

Logographic writing systems

➤ Logographs (also called Logograms):

    ➤ Pictographs (Pictograms): originally pictures of things, now stylized and simplified.

    ➤ Ideographs (Ideograms): representations of abstract ideas

    ➤ Compounds: combinations of two or more logographs.

    ➤ Semantic-phonetic compounds: symbols with a meaning element (hints at meaning) and a phonetic element (hints at pronunciation).

➤ Examples: Chinese (Japanese, Vietnamese), Mayan, Ancient Egyptian

Development of Chinese character horse

Type of Characters		Descriptions
金文	🐟	Bronze script Jin wen 15th - 11th centuries B.C.E.
甲骨文		Oracle-bone script Jia gu wen 12th - 11th centuries B.C.E.
大篆		Large-seal script da chuan c. 8th century B.C.E.
小篆		Small-seal script xiao chuan 2nd century B.C.E.
隸書	馬	Clerical script li shu 2nd century C.E.
楷書	馬	Standard script k'ai shu since c. 4th century C.E.
行書	馬	Running script Xing shu since c. 4th century C.E.
草書	马	Cursive script Chao shu since c. 4th century C.E.

➢ Pictographs

Pictographs

女	子	口	日	月	山	川	豕	目	心	雨	田	木	龜
woman	child	mouth	sun	moon	mountain	river	pig	eye	heart	rain	field	tree	turtle

➢ Ideographs

Ideographs

一	二	三	上	下	中	力	凸	凹
one	two	three	above	below	middle	stength (plough)	convex	concave

➢ Compounds of Pictographs/Ideographs

Compound Pictographs / Compound Ideographs

好	安	明	家	思	牢	雷	男
good (woman + child)	peaceful (woman under a roof)	bright (sun + moon)	home/family (pig under a roof)	thought (heart + field)	prison (cow under a roof)	thunder (rain cloud over a field)	man/male (field + strength)

Semantic-phonetic compounds

Semantic-phonetic compounds

phonetic component

	古 gǔ	扁 biǎn	敖 aó	旁 páng	堯 yǎo
人 (person)	估 gū (to guess)	偏 piān (biased)	傲 ào (proud)	傍 bāng (beside)	僥 jiǎo (lucky)
言 (words)	詁 gū (commentaries)	論 pián (to quibble)	警 áo (to slander)	謗 bàng (to libel)	譊 náo (to argue)
虫 (insect)	蛄 gū (mole cricket)	蝙 biān (bat)	螯 áo ([crab's] nippers)	螃 páng (crab)	蟯 ráo (worm)
金 (metal)	鈷 gū (cobalt)		鏊 áo (griddle)	鎊 bàng (pound sterling)	鐃 náo (cymbals)

semantic component (radical)

97% of Chinese characters are phonetic compounds (Sproat 2010)! Other estimates are lower (Wang, pc 2013)

An example from Ancient Egyptian
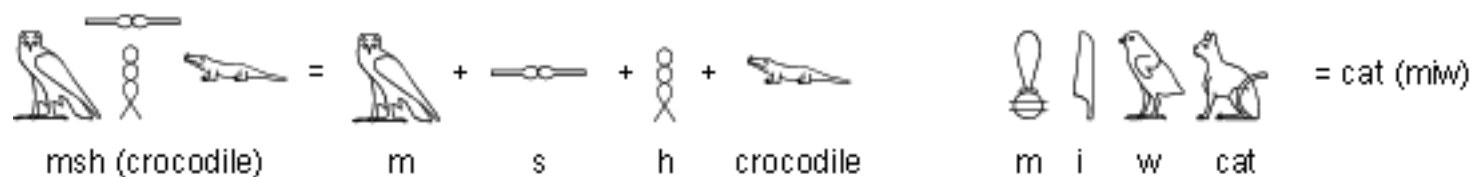
Redundant, with both pronunciation and meaning! Sometimes only one would be used, sometimes the other, sometimes both. Gradually pronunciation won out and the characters were simplified and became Hieratic and then Demotic.

Relating writing systems to languages

➢ There is never a simple correspondence between a writing system and a language.

➢ For example, English uses the Roman alphabet, but Arabic numerals (e.g., 3 and 4 instead of III and IV).

➢ Even when a new alphabet is designed, pronunciation changes.

Comparison of writing systems

The pros and cons of each type of system depend on a variety of factors:

➢ Accuracy: Can every word be written down accurately?

➢ Learnability: How long does it take to learn the system?

➢ Cognitive ability: Are some systems unnatural? (e.g. Does dyslexia show that alphabets are unnatural?)

➢ Language-particular differences: English has thousands of possible syllables; Japanese has very few in comparison (52, modulo vowel length)

➢ Connection to history/culture: Is there meaning in the system beyond its use as a writing system?

➢ Some languages have changed scripts (sometimes more than once):

➢ Malay: pallava (Brahmic), jawi (Arabic), alphabet
➢ Korean: hanja (Chinese), hangul
➢ Mongolian: Hudum Mongol bichig (Old Uygar, from Aramaic), Cyrillic

➢ Many languages have had orthographic reforms

➢ American Spelling (English)
   ∗ *authour* → *author*
➢ Japanese post-war reform (actually started pre-war)
   ∗ 蝶てふ *tefu* → ちょう *chou* "chi-xyo-u" "butterfly"
   ∗ 居るゐる *wiru* → いる *iru* "be"
➢ New Rumi Spelling/Ejaan Yang Disempurnakan (Malay/Indonesian) in 1972
   ∗ *di-buat* → *dibuat*, *di-rumah* → *di rumah*
   ∗ *ch/tj* /tʃ/ → *c*: *tjicak/chicak* → *cicak*
   ∗ *njonja/nyonya* → *nyonya*

Encoding written language

➤ Information on a computer is stored in **bits**.

➤ A bit is either on (= 1, yes) or off (= 0, no).

➤ A list of 8 bits makes up a byte, e.g., 01001010

➤ Just like with the base 10 numbers we're used to, the order of the bits in a byte matters:

  ➤ Big Endian: most important bit is leftmost (the standard way of doing things)
  ➤ Little Endian: most important bit is rightmost (only used on Intel machines)

How much information in a byte?

➤ Every bit encodes two states (1 or 0)

➤ n bits encodes 2^n states

  ➤ $2 \times 2 \times 2 \times 2 \ldots n$ times

➤ So 8 bits encodes 2^8 or 256 things

An encoding standard: ASCII

➢ With 256 possible characters, we can store:

  ➢ every single letter used in English,
  ➢ plus all the things like commas, periods, space bar, percent sign (%), back space, and so on.

➢ **ASCII** = the American Standard Code for Information Interchange

  ➢ 7-bit code for storing English text
  ➢ 7 bits = 128 possible characters.
  ➢ The numeric order reflects alphabetic ordering.

The ASCII chart

Codes 1–31 are used for control characters (backspace, return, tab, …).

032	␣	048	0	064	@	080	P	096	`	112	p
033	!	049	1	065	A	081	Q	097	a	113	q
034	"	050	2	066	B	082	R	098	b	114	r
035	#	051	3	067	C	083	S	099	c	115	s
036	$	052	4	068	D	084	T	100	d	116	t
037	%	053	5	069	E	085	U	101	e	117	u
038	&	054	6	070	F	086	V	102	f	118	v
039	`	055	7	071	G	087	W	103	g	119	w
040	(	056	8	072	H	088	X	104	h	120	x
041	)	057	9	073	I	089	Y	105	i	121	y
042	*	058	:	074	J	090	Z	106	j	122	z
043	+	059	;	075	K	091	[	107	k	123	{
044	'	060	<	076	L	092	\	108	l	124	_
045	–	061	=	077	M	093	]	109	m	125	}
046	.	062	>	078	N	094	∧	110	n	126	∼
047	/	063	?	079	O	095	_	109	o	127	DEL

What if 127 characters isn't enough?

➢ Local Variants

[092] Japanese ASCII: Yen (¥) instead of backslash (\)
[035] UK ASCII: Pounds Sterling (£) instead of hash (#).

➢ But what if you need more letters?

    ➢ Transliteration
    ➢ Multi-byte encodings

Transliteration

➤ Use ASCII, and fake the missing letters

    ➤ ue for ü, oe for ö, …

➤ Volapuk replaces Cyrillic letters with Latin ones in order to look the same as typed or handwritten Cyrillic letters.

    1. Replace "the same" letters: a, e, K, M, T, o, y.
    2. Replace similar-looking letters: Γ with 2 (handwritten resemblance) or r,…
    3. Replace all other non-obvious hard-to-represent characters; there are many options for each letter: Φ with qp or 0. The choice for each letter depends on the preferences of the individual user.

➤ These transliterations are hard to read, but better than nothing

Different coding systems

➤ Extended ASCII (use 256 characters)

➤ Other encodings

  ➤ ISO 8859-1: includes extra letters for French, German, Spanish, …
  ➤ ISO 8859-5: Cyrillic alphabet
  ➤ ISO 8859-7: Greek alphabet
  ➤ ISO 8859-8: Hebrew alphabet

➤ But you can only have one encoding at a time!

  ➤ You can't have both Greek and Russian

➤ 256 characters is not enough for many languages

Multi-byte Encodings

➤ Use more bytes

➤ EUC-JP (Extended Unix Code Japanese)

  ➤ An ASCII character is represented by one byte, with the first bit 0.
  ➤ A character from JIS-X-0208 (code set 1) is represented by two bytes, both with the first bit 1.
    * This includes Hiragana, Katakana and most Chinese Characters.
  ➤ A character from JIS-X-0212 (code set 3) is represented by three bytes, the first being 0x8F, and the second two both with the first bit 1.
    * This includes many more Chinese characters.

  This encoding scheme allows the easy mixing of 7-bit ASCII and 8-bit Japanese.

Example of EUC-JP

犬	は	d	o	g	だ	。	EOF
b8a4	a4cf	64	6f	67	a4c0	a1a3	0a

➤ Written in hexadecimal: 0123456789ABCDEF

  ➤ 0 = 0000 = 0 (0 + 0 + 0 + 0)
  ➤ 1 = 0001 = 1 (0 + 0 + 0 + 1)
  ➤ 2 = 0010 = 2 (0 + 0 + 2 + 0)
  ➤ 4 = 0100 = 4 (0 + 4 + 0 + 0)
  ➤ 8 = 1000 = 8 (8 + 0 + 0 + 0)

  …

  ➤ A = 1010 = 10 (8 + 0 + 2 + 0)

  …

  ➤ E = 1110 = 14 (8 + 4 + 2 + 0)
  ➤ F = 1111 = 15 (8 + 4 + 2 + 1)

➤ Bit one $= 1 \Rightarrow\; > 8$

➤ So b8a4 is 1011 1000 1010 0100

Problems with stateless encodings

➤ Shift-JIS takes a different approach

➤ each character is two bytes, so you can use the 8th bit

➤ the trouble is you have to know where you are, …

    ➤ Consider the following:

剣	道		*kendo* "kendo"
8C95	93B9		
白	血	病	
9492	8C8C	9561	*hakketsubyō* "leukemia"

    ➤ 剣 8C95 matches across character boundaries
       ∗ but you don't want to match it here

➤ When you delete a character, you need to know how many bytes it is

Still more problems

➤ EUC-JP is stateful so it can't fit all characters

  ➤ using one bit to show state, so only: $2^{14} = 16{,}384$

➤ You need to know what the encoding is:

  ➤ "æ–‡å—åŒ–ã[] '" (文字化け)

Much more in:
Lunde, K. (1999). *CJKV Information Processing*. O'Reilly, Sebastopol, CA

Unicode

➤ Unicode solves many of these problems

"Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language." (`www.unicode.org`)

How big is Unicode?

➢ Version 3.2 (2002) has codes for 95,221 characters from alphabets, syllabaries and logographic systems.

➢ Uses 32 bits (4 bytes)
Can represent $2^{32} = 4, 294, 967, 296$ characters.

➢ 4 billion possibilities for each character?

➢ That takes a lot of space on the computer!

➢ Four times as much as ASCII

➢ Unicode 11.0 (2018) contains 137,439 characters covering 146 modern and historic scripts, as well as multiple symbol sets and emoji.

➢ Version 15.1 (2023) defines 149,813 character covering 161 scripts

Compact encoding of Unicode characters

➢ UTF-32 (32 bits): direct representation

➢ UTF-16 (16 bits): $2^{16} = 65{,}536$ (subset!)

➢ UTF-8 (variable width encoding)

U+0000-U+007F	0xxxxxxx				ASCII
U+0080-U+07FF	110yyyxx	10xxxxxx			Alphabets/Syllabaries
U+0800-U+FFFF	1110yyyy	10yyyyxx	10xxxxxx		Logographs
U+10000-U+10FFFF	11110zzz	10zzyyyy	10yyyyxx	10xxxxxx	Room to expand

➢ First byte says how many will follow

Nice consequence: ASCII text is in a valid UTF-8 encoding.

How do we type everything in?

➢ Use a keyboard tailored to your specific language

➢ e.g. Highly noticeable how much slower your English typing is when using a Danish-designed keyboard.

➢ Use a processor that allows you to switch between different character systems.

   ➢ e.g. Type in Cyrillic characters on your English keyboard.

➢ Use combinations of characters.

   ➢ e.g. An e followed by an ' might result in an é.

➢ Pick and choose from a table of characters.

Unwritten languages

Many languages have never been written down. Of the 6,900 spoken languages, approximately 3,000 have never been written down.

Some examples:

➤ Salar, a Turkic language in China.

➤ Gugu Badhun, a language in Australia.

➤ Southeastern Pomo, a language in California

Ongoing work in adding alphabets, often by Bible translators and linguists!

Redundancy of Representation

➤ You can remove a lot of information and still understand

  ➤ For example, with no vowels, spaces or segmentation
  ➤ F C T S R S T R N G R T H N F C T N

➤ It is much easier if you know the meaning

➤ Redundancy is important if there is noise

➤ There is normally a lot of noise, so all natural languages are redundant

Efficient Representation

➤ Language is also efficient in its coding

➤ Consider the most common 20 words of English:
the, of, and, a, to, in, is, you, that, it, he, was, for, on, are, as, with, his, they, I

➤ They are all short!

➤ Frequent expressions are shortened
today not *on this day*

➤ This makes the overall text shorter

Playing with Writing: Acrostics

To the Members of the California State Assembly:
I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely, Arnold Schwarzenegger

Wired (2009) "SCHWARZENEGGER FLIPS OFF LAWMAKERS IN HIDDEN MESSAGE" https://www.wired.com/2009/10/schwarzenegger/

Playing with Writing: Acrostics

To the Members of the California State Assembly:
I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many
unnecessary bills come to me for consideration. Water reform, prison reform, and health
care are major issues my Administration has brought to the table, but the Legislature just
kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians
overwhelmingly deserve. In light of this, and after careful consideration, I believe it is
unnecessary to sign this measure at this time.

Sincerely, Arnold Schwarzenegger

➢ 目田氏王 *mùtián shīwáng* "eye-field clan-king"

Playing with Writing: Graphemic Puns

➢ 目田氏王 *mùtián shīwáng* "eye-field clan-king"

➢ 自由民主 *zìyóu mínzhǔ* "freedom democracy"

　➢ Just like freedom and democracy but missing a bit
　➢ Not caught by censors (at first)

Victor Mair (2010) "Decapitated Democracy, Headless Liberty"

Language Log http://languagelog.ldc.upenn.edu/nll/?p=2614

Speech vs Writing (1): Time-bound

➤ **Speech**

  ➤ time-bound
  ➤ dynamic, transient
  ➤ normally direct between a speaker and a known addressee

➤ **Writing**

  ➤ space-bound
  ➤ static, permanent
  ➤ normally indirect with the addressee unknown

Summary of Table 2.1 (pp 28–30) in Crystal, D. (2006). *Language and the Internet.*
Cambridge University Press, 2nd edition. (conversation vs books)

Speech vs Writing (2): Spontaneous

➤ **Speech**

➤ no lag between production and reception
➤ hard to plan complex constructions
⇒ repetitions, rephrasing, comments clauses
➤ Sentence boundaries often unclear

➤ **Writing**

➤ lag between production and reception
➤ readers can reread and analyse in depth
⇒ careful organization and compact expressions
➤ Sentence (and paragraph, …) boundaries are clear   (English!)

Speech vs Writing (3): Face-to-Face

➢ **Speech**

  ➢ Extralinguistic cues are common (facial expressions, gestures)
  ➢ Immediate feedback (back channel)
  ➢ Deictic expressions are common (referring to the situation)
    that one, you, now, over there

➢ **Writing**

  ➢ Different extralinguistic possibilities (fonts, color, pictures)
  ➢ No immediate feedback
  ➢ Fewer deictic expressions

Speech vs Writing (4): Loosely Structured

➢ **Speech**

    ➢ Contractions are common: *isn't, he's*
    ➢ Long coordinate sentences are common
    ➢ informal vocabulary: *thingamajig, whatsit*
    ➢ obscenity more common

➢ **Writing**

    ➢ Subordination more common (relative clauses)
    ➢ Longer sentences (can be multipage)
    ➢ Some items rarely pronounced: $H(\mathrm{p}) = -\sum_{x \in X} \mathrm{p}(x) \log_2 \mathrm{p}(x)$

Speech vs Writing (5): Socially Interactive

➢ **Speech**

  ➢ Well suited to social functions
  * greetings
  * maintaining social relationships
  * expression attitudes and opinions
  Much use of prosody and non-verbal features

➢ **Writing**

  ➢ Suited to recording facts and communicating ideas
  ➢ Easier to scan
  ➢ Tables demonstrate relations between things
  ➢ Text can be read at one's own pace

Speech vs Writing (6): Immediately Revisable

➢ **Speech**

> ➢ You can rephrase at once, based on feedback
> ➢ Errors once spoken can't be withdrawn
> ➢ Interruptions and overlap is common

➢ **Writing**

> ➢ You can remove errors without the speaker ever seeing them
> ➢ Once published errors can only be withdrawn through revisions
> ➢ Interruptions are not visible

Speech vs Writing (7): Prosodically Rich

➢ **Speech**

   ➢ Prosody
      intonation; loudness; tempo; rhythm, pause tone of voice

➢ **Writing**

   ➢ Pages, lines, capitalization, spatial organization
   ➢ Punctuation (?!.)
   ➢ **Fonts**, CAPITALIZATION, *style*
   ➢ Tables, graphs, formulae

Comparison of speed for different modalities

Speed in words per minute (one word is 6 characters)
(English, computer science students, various studies)

Activity	Speed (wpm)	Comments
Reading	300	200 (proof reading)
Writing	31	21 (composing)
Speaking	150	
Hearing	150	210 (speeded up)
Typing	33	19 (composing)

➤ Reading >> Speaking/Hearing >> Typing

⇒ Speech for input
⇒ Text for output

Does anyone know of work on languages other than English?

Summary

➢ There are many ways to represent text

➢ Some are easier to encode than others

➢ Efficient representation is not always the goal

➢ Speech can be very different from text

Acknowledgments

➢ Many slides taken from Marcus Dickinson

➢ Much of the information on writing systems and the graphics used are taken from the great site `http://www.omniglot.com`

Just for interest

A writing system with an unusual realization: Braille

➢ Tactile

    ➢ Braille is a writing system that makes it possible to read and write through touch; primarily used by the (partially) blind.

    ➢ It uses patterns of raised dots arranged in cells of up to six dots in a 3×2 configuration.

        ∗ How many bits is this?

        ∗ How many characters can be represented?

    ➢ Each pattern represents a character, but some frequent words and letter combinations have their own pattern.

Braille alphabet

Basic letters

a	b	c	d	e	f	g	h	i	j	k	l	m
n	o	p	q	r	s	t	u	v	w	x	y	z

Accented letters

à	â	ä/æ	è	é	ê	ë	ì	î
ï	ò	ô	ö/œ	ù	û	ü	ç	

Words and abreviations

a	but	can	do	every	from	go	have	just	knowledge	like	more	not
people	quite	rather	so	that	us	very	will	it	you	as	and	for
of	the	with	child/ch	gh	shall/sh	this/th	which/wh	ed	er	out/ou	ow	bb
cc	dd	en	gg; were	in	st	ing	ar					

Braille terminals (refreshable Braille displays) push the pins up in real time

Braille Secrets

(Cartoon from http://xkcd.com/315/)