Writing Systems, Transliteration and Decipherment

Kevin Knight
University of Southern California
Information Sciences Institute
Richard Sproat
Oregon Health & Science University
Center for Spoken Language Understanding

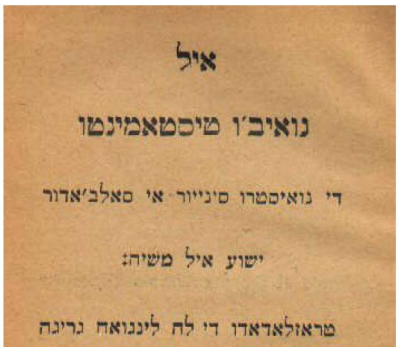
Overview

- An overview of writing systems
- Transcription/transliteration between scripts
- Traditional and automatic approaches to decipherment

Part I
Writing Systems and Encodings

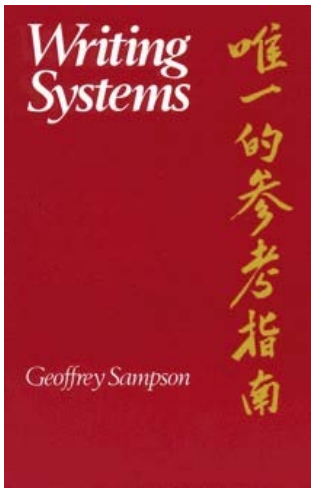
Some terminology

- A *script* is a set of symbols
- A *writing system* is a *script* paired with a *language.*



What could writing systems represent?

- In principle any linguistic level



"My dog likes avocados"

maɪ dɔg laɪks ævəkɑdoz

(maɪ) (dɔg) (laɪks) (æ)(və)(kɑ)(doz)

[(æ)(və)][(kɑ)(doz)]

maɪ dɔg laɪk+s ævəkɑdo+z



What do writing systems actually represent?

- Phonological information:
  - Segmental systems:
    - Alphabets
    - *Abjads*
    - *Alphasyllabaries*
  - Syllables (but *full* syllabaries are *rare*)
- Words in partially *logographic* systems
- *Some* semantic information:
  - Ancient writing systems like Sumerian, Egyptian, Chinese, Mayan
- **But no full writing system gets by without some representation of sound**

Roadmap

- Look at how Chinese writing works: Chinese is the only “ancient” writing system in current use, and in many ways it represents how all writing systems used to operate.
- Detour slightly into “semantic-only” or “logographic” writing.
- Survey a range of options for phonological encoding

The “six writings”

- *xiàngxíng* simple pictograms
  - ‘person’, ‘wood’, ‘turtle’
- *zhǐshì* indicators
  - ‘above’, ‘below’
- *huìyì* meaning compound
  - ‘bright’ (SUN+MOON)
- *xíngshēng* phonetic compounds
  - ‘oak’ (TREE+*xiàng*), ‘duck’ (BIRD+*jiǎ*)
- *zhuǎnzhù* ‘redirected characters’
  - ‘trust’ (PERSON+WORD)
- *jiǎjiè* ‘false borrowings’ (rebuses)
  - ‘come’ (from an old pictograph for ‘wheat’)

Xíngshēng characters

95% of Chinese Characters ever invented consist of a *semantic* and a *phonetic* component

鯉 = 魚 → · 里

lǐ (‘carp’) = FISH · LI

鴨 = 鳥 ← · 甲

yā (‘duck’) = BIRD · JIA

草 = 艸 ↓ · 早

cǎo (‘grass’) = VEGETATION · ZAO

志 = 心 ↑ · 士

zhì (‘will, goal’) = HEART · SHI

國 = 口 ⊙ 或

guó (‘country’) = ENCLOSURE · HUO

General operation: SEMANTIC · PHONETIC

General operation: SEMANTIC · PHONETIC

鯉 = 魚 → · 里
lǐ ('carp') = FISH · LI

鴨 = 鳥 ← · 甲
yā ('duck') = BIRD · JIA

草 = 艸 ↓ · 早
cǎo ('grass') = VEGETATION · ZAO

志 = 心 ↑ · 士
zhì ('will, goal') = HEART · SHI

國 = 口 ⊙ 或
guó ('country') = ENCLOSURE · HUO

General operation: SEMANTIC · PHONETIC

鯉 = 魚 → · 里
lǐ ('carp') = FISH · LI

鴨 = 鳥 ← · 甲
yā ('duck') = BIRD · JIA

草 = 艸 ↓ · 早
cǎo ('grass') = VEGETATION · ZAO

志 = 心 ↑ · 士
zhì ('will, goal') = HEART · SHI

國 = 口 ⊙ 或
guó ('country') = ENCLOSURE · HUO

General operation: SEMANTIC · PHONETIC

鯉 = 魚 → · 里
lǐ ('carp') = FISH · LI

鴨 = 鳥 ← · 甲
yā ('duck') = BIRD · JIA

草 = 艸 ↓ · 早
cǎo ('grass') = VEGETATION · ZAO

志 = 心 ↑ · 士
zhì ('will, goal') = HEART · SHI

國 = 口 ⊙ 或
guó ('country') = ENCLOSURE · HUO

General operation: SEMANTIC · PHONETIC

鯉 = 魚 → · 里
lǐ ('carp') = FISH · LI

鴨 = 鳥 ← · 甲
yā ('duck') = BIRD · JIA

草 = 艸 ↓ · 早
cǎo ('grass') = VEGETATION · ZAO

志 = 心 ↑ · 士
zhì ('will, goal') = HEART · SHI

國 = 口 ⊙ 或
guó ('country') = ENCLOSURE · HUO

A generalization of *huìyì*: Japanese *kokuji* (国字)

Alex. #	Kokuji	Analysis	(Phonetic)	Kun	(*on*)	Gloss
10	働	<PERSON+MOVE>		*hataraki*	*dō*	'effort'
12	凪	<WIND+STOP>		*nagi*		'lull, calm'
33	峠	<MOUNTAIN+UP+DOWN>		*touge*		'mountain pass'
37	悚	<HEART+FOREVER>		*kore*		'endure'
74	毟	<FEW+HAIR>		*mushi*		'pluck'
124	聢	<EAR+CERTAIN>		*shika*		'clearly'
160	躾	<BODY+BEAUTIFUL>		*shitsuke*		'upbringing'
198	颪	<DOWN+WIND>		*oroshi*		'mountain wind'
240	鴫	<FIELD+BIRD>		*shigi*		'snipe'
249	嬶	<FEMALE+NOSE>		*kakā*		'wife'
138	薙	<GRASS+ZA>	座 *za* (*on*)		*goza*	'matting'
51	柾	<TREE+MASA>	正 *masa* (*kun*)	*masa*		'straight grain'
147	袴	<CLOTHING+YUKI>	行 *yuki* (*kun*)	*yuki*		'sleeve length'

Japanese logography

- Japanese writing has three subsystems
  - Two *kana* syllabaries, which we'll look at later
  - Chinese characters – *kanji* which usually have two kinds of readings:
    - Sino-Japanese (*on* 'sound') readings: often a given character will have several of these
    - Native Japanese (*kunyomi*) readings

		鯉
'mountain'	'island'	**'carp**
***on:* san**	*on:* too	***on:* ri**
***kun:* yama**	*kun:* shima	***kun:* koi**

A generalization of *xíngshēng*: Vietnamese *Chữ Nôm* (____)

Semantic-phonetic constructions in other ancient scripts

Egyptian

Su

[DIV]

Urim

w3ḥ (verb 3-lit.) "set, place; add; stop; remain, last"; *w3ḥ* (adjective) "lasting"; *w3ḥ jb* "be patient" (literally, "lasting/set of heart")

w3ḥyt (noun) "abundance (of grain)"

w3ḫj (noun) "columned hall" (literally, "marsh" of papyrus and lotus columns)

w3s (noun) "dominion"

w3st (noun) "Thebes" (nome and city)

w3s (noun) "ruin" (infinitive of 4ae-inf. verb *wˁsj* "fall into ruin")

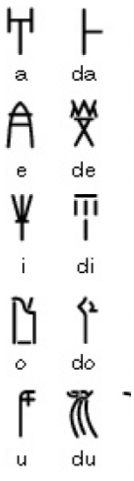
w3gj (verb 4ae-inf.) "make festival"

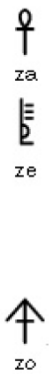
Ba Balam Ma

Syllabaries

- Syllables are often considered more “natural” representations in contrast to phonemes. E.g:
  - *“investigations of language use suggest that many speakers do not divide words into phonological segments unless they have received explicit instruction in such segmentation comparable to that involved in teaching an alphabetic writing system”* [Faber, Alice. 1992. “Phonemic segmentation as epiphenomenon. evidence from the history of alphabetic writing.” In Pamela Downing, Susan Lima, and Michael Noonan, eds, *The Linguistics of Literacy*. John Benjamins, Amsterdam, pages 111--34.]
- Syllabaries have been invented many times (true); the alphabet was only invented once (not so clearly true)
- But: very few systems exist that have a separate symbol for every syllable of the language:
  - Most are *defective* or at least partly segmental

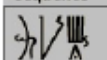
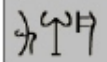
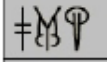
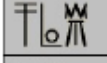
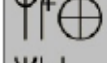
Linear B (ca 1600-1100 BC)



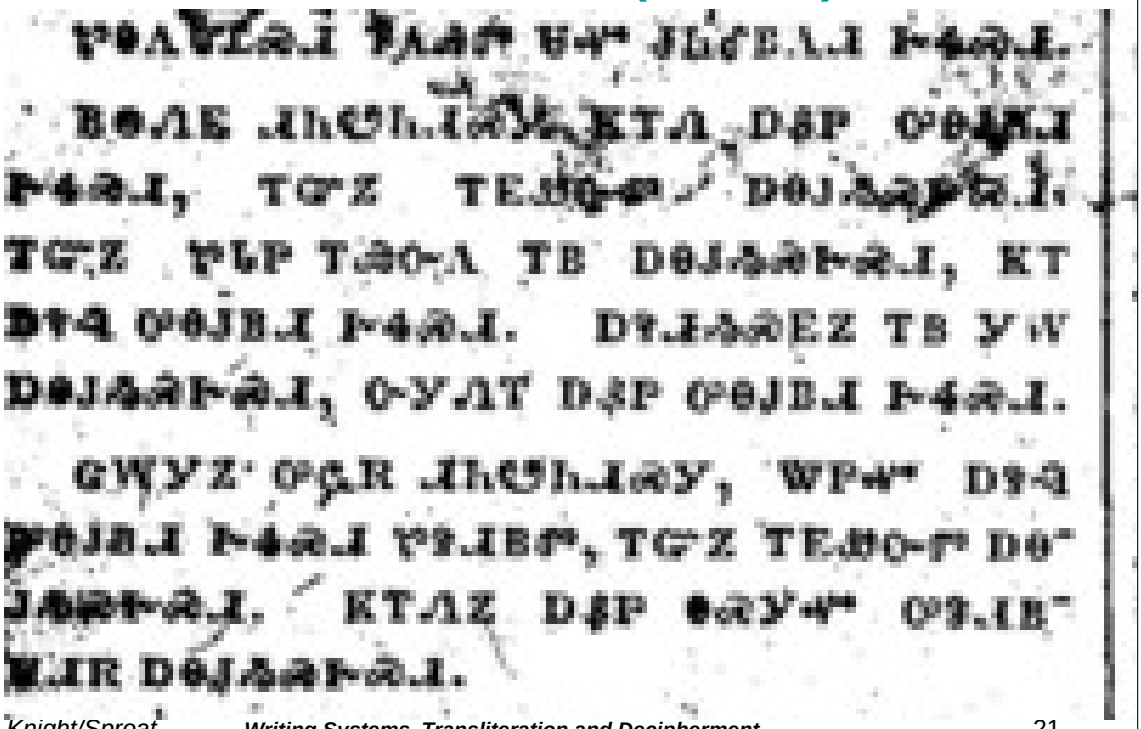


Derived from an earlier script (Linear A), which was used to write an unknown language (Minoan)

Linear B

sign sequence	trans-literation	Mycenaean Greek	Classic Greek	word meaning
	ku-mi-no	*kuminon	kuminon	cumin
	ku-na-ja	*gunaia	gune	woman (*gynecology*)
	ku-ru-so	*khrusos	khrusos	gold (*chrysanthemum*)
	pa-te	*pater	pater	father
	pa-ma-ko	*pharmakon	pharmakon	medicine (*pharmacy*)
	to-so	*toso	tosos	so many
	to-ra-ke	*thorakes	thorax	thorax
	qo-u-	*gwou-	bou-	cow
	i-qo	*hikkwoi	hippos	horse
	re-u-ka	*leuka	leukos	white (*leukemia*)
	re-a	*rea	rhis, rhino-	nose (*rhinoplasty*)

Cherokee (1821)



Kana

万葉仮名 *man'yōgana*

Syl.	Kata.	Char.	Mand.	Mid. Chin.	Hira.	Char.	Mand.	Mid. Chin.
a	ア	阿	ā	ʔa	あ	安	ān	ʔan
i	イ	伊	yī	ʔjij	い	以	yǐ	yiX
u	ウ	宇	yǔ	hjuX	う			
e	エ	江			え	衣	yī	ʔjij
o	オ	於	yú	ʔjo	お			
ka	カ	加	jiā	kæ	か			
ki	キ	幾	jī	gɨj	き			
ku	ク	久	jiǔ	kjuwX	く			
ke	ケ	介	jiè	kɛjH	け	計	jì	kejH
ko	コ	己	jǐ	kiX	こ			
sa	サ	散	sàn	sanH	さ	左	zuǒ	tsaX
shi	シ	之	zhī	tsyi	し			
su	ス	須	xū	sju	す	寸	cùn	tshwonH
se	セ	世	shì	syejH	せ			
so	ソ	曾	zēng	tsong	そ			
wo	ヲ	平	píng	bjæng	を	遠	yuǎn	hjwenH
N	ン	无	wú	mju	ん			

テネシー *teneshii*

Yi

Figure 2.13: Yi syllabary. http://www.omniglot.com/images/writing/yi_syl.gif

Segmental writing

- Somewhere around 3000 BC, the Egyptians developed a mixed writing system whose phonographic component was essentially *consonantal* – hence segmental
- One hypothesis as to why they did this is that Egyptian – like distantly related Semitic – had a *root and pattern* type morphology.
  - Vowel changes indicated morphosyntactic differences; the consonantal root remained constant
  - Thus a spelling that reflected only consonants would have a constant appearance across related words

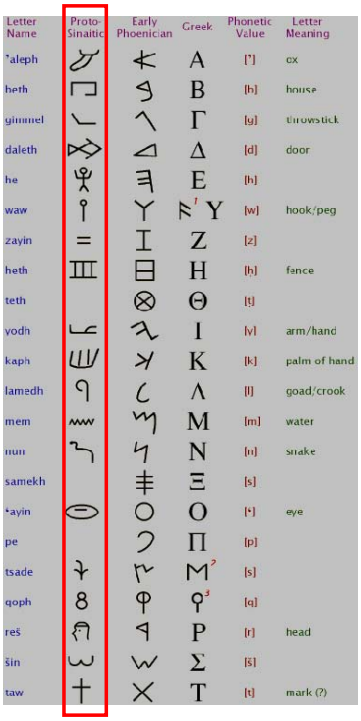
Egyptian consonantal symbols

Figure 2.16: Egyptian triconsonantal symbols, from `http://www.omniglot.com/images/writing/hieroglyphs_trilat.gif`.

Proto-Sinaitic (aka Proto-Canaanite) script

- Somewhere around 2000 BC, Semitic speakers living in Sinai, apparently influenced by Egyptian, simplified the system and devised a consonantal alphabet
- This was a *completely* consonantal system:
  - No *matres lectionis* – using consonantal symbols to represent long vowels – as in later Semitic scripts
- Phoenician (and other Semitic scripts) evolved from this script

Proto-Sinaitic script



Later Semitic scripts: vowel diacritics

בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ

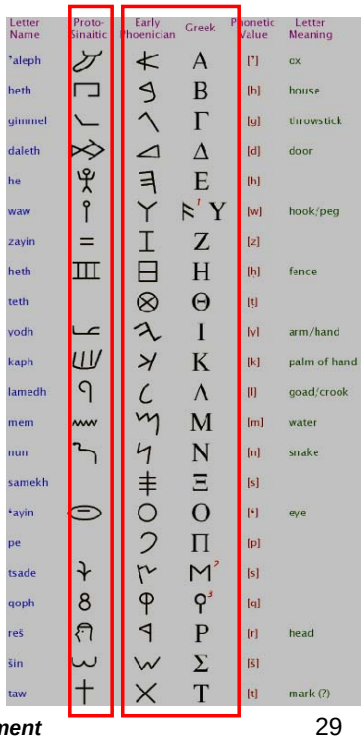
بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ

ܒܫܡ ܐܠܗܐ ܪܚܡܢܐ ܪܚܝܡܐ.

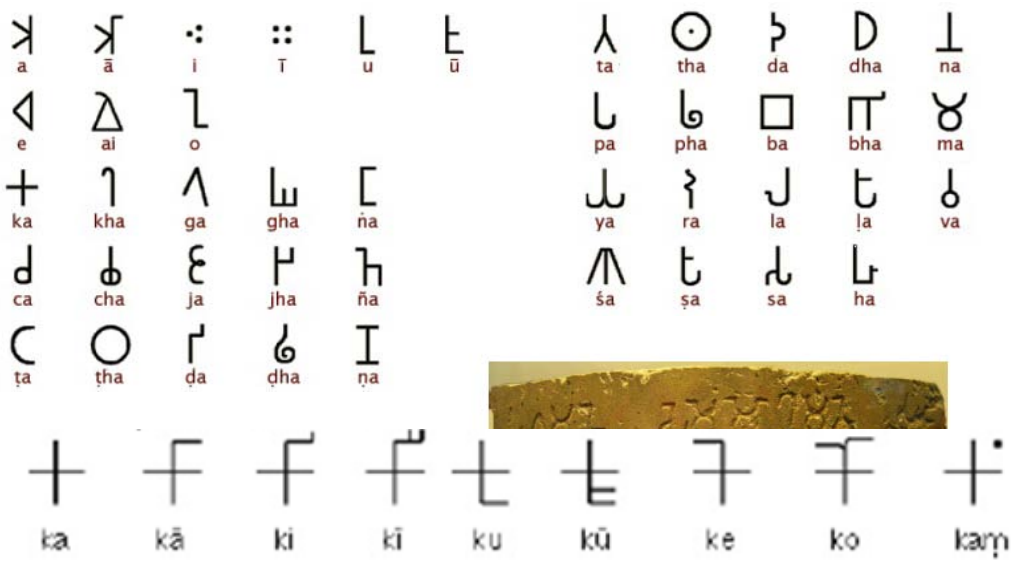
בּשְׁמָא דַּאֱלָהָא רַחֲמָנָא רַחִימָא.

The evolution of Greek writing

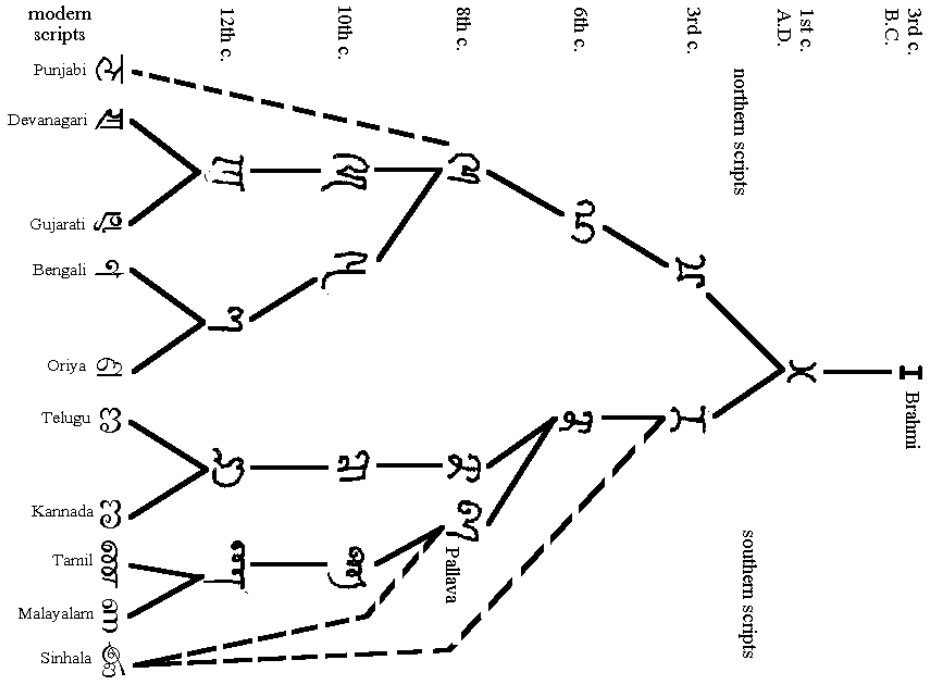
- Greek developed from Phoenician
- Vowel symbols developed by reinterpreting – or maybe *mis*interpreting – Phoenician consonant symbols
- The alphabet is often described as only having been invented once.
  - But that's not really true: the Brahmi and Ethiopic *alphasyllabaries* developed apparently independently, from Semitic



Alphasyllabaries: Brahmi (ca 5th century BC)



Some Brahmi-derived scripts



Basic design of Brahmi-derived alphasyllabaries

- Every consonant has an *inherent* vowel
  - This may be canceled by an explicit *cancellation sign* (*virama* in Devanagari, *pulli* in Tamil)
  - Or replaced by an explicit vowel diacritic
- In many scripts consonant groups are written with some consonants subordinate to or ligatured with others
- In most scripts of India vowels have separate full and diacritic forms:
  - Diacritic forms are written after consonants
  - Full forms are written syllable or word initially
  - In most Southeast Asian scripts (Thai, Lao, Khmer …) this method is replace by one where *all* vowels are diacritic, and syllables with open onsets have a special empty onset sign. (We will see this method used again in another script.)

Devanagari vowels

Catenator	Full form		Diacritic form	
Null	अ	<a>	क	<ka>
→	आ	<aa>	का	<kaa>
	ओ	<o>	को	<ko>
	औ	<au>	कौ	<kau>
	ई	<ii>	की	<kii>
↑	ए	<e>	के	<ke>
	ऐ	<ai>	कै	<kai>
↓	उ	<u>	कु	<ku>
	ऊ	<uu>	कू	<kuu>
	ऋ	<ri>	कृ	<kri>
←	इ	<i>	कि	<ki>

Inherent vowel

Kannada diacritic vowels

Null	ಕ	<ka>
→	ಕು	<ku>
	ಕಾ	<kaa>
	ಕೂ	<kuu>
↑	ಕಿ	<ki>
	ಕೆ	<ke>
	ಕೌ	<kau>
↓	ಕೃ	<kri>
	ಕೄ	<krii>
{i (↑), LONG (→)}	ಕೀ	<kii>
{e (↑), uu (→)}		<ko>
{e (↑), ai (↓)}	ಕೈ	<kai>
{e (↑), LONG (→)}	ಕೇ	<kee>
{e (↑), uu (→), LONG (→)}	ಕೋ	<koo>

<koo> ಕೋ

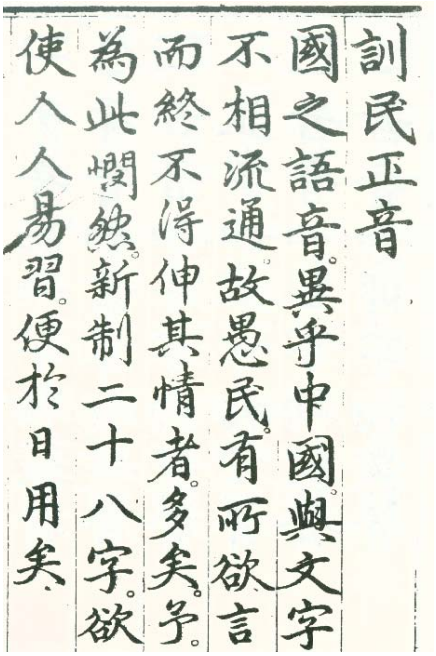
<lakṣmiiṣa>

Another alphasyllabary: Ethiopic (Ge'ez) (4th century AD)

æ	u:	i:	a:	e:	ə	o:
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ

may	m	መ	መ	ሚ	ማ	ሜ	ም	ሞ	ሟ	ሸ
säwt	ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ	
re's	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	ሯ	ሪ
śat	ś	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ	
kaf	k	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቋ	
bet	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ	
täwe	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ	ቷ	
ḫarm	ḫ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኋ	
nähas	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ	ኗ	
'älf	'	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	ኧ	

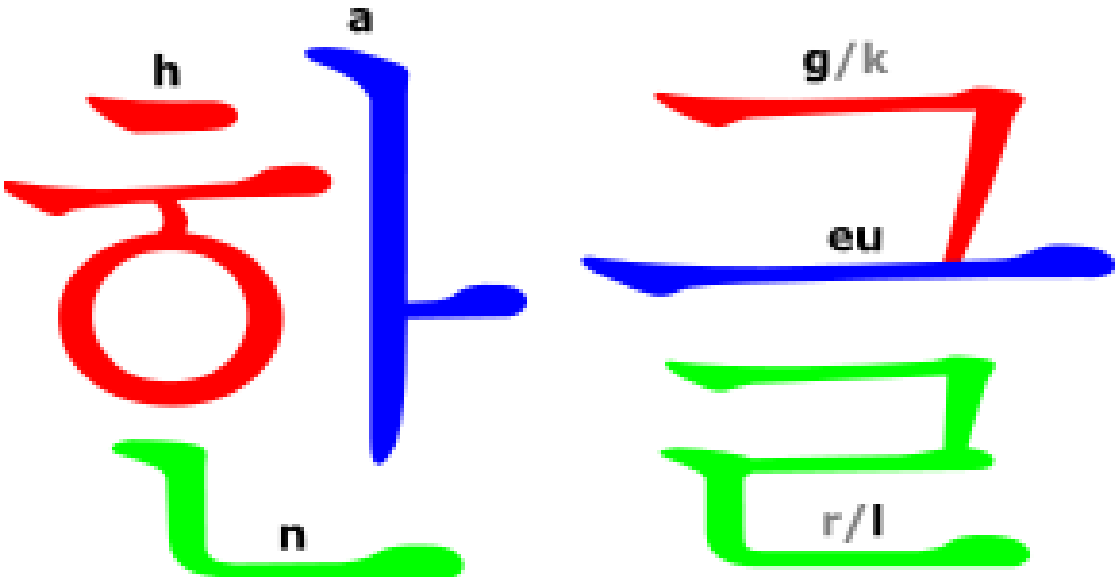
“Correct sounds for instructing the people”
(훈민정음)
The origin of Korean Writing


訓民正音
國之語音。異乎中國。與文字
不相流通。故愚民。有所欲言
而終不得伸其情者。多矣。予
爲此憫然。新制二十八字。欲
使人人易習。便於日用矣。

“The speech of our country differs from that of China, and the Chinese characters do not match it well. So the simple folk, if they want to communicate, often cannot do so. This has saddened me, and thus I have created twenty eight letters. I wish that people should learn the letters so that they can conveniently use them every day.”

King Sejong the Great
(Chosun Dynasty, 1446)

Hangul symbols



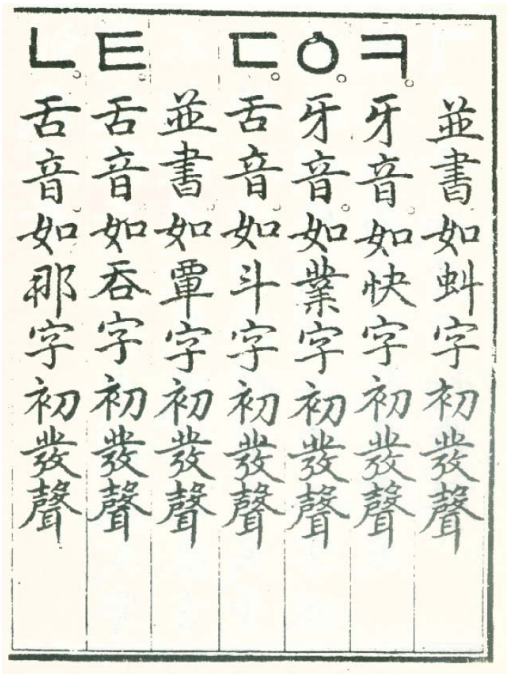
Design principles of Hangul

- For consonants based on the position of articulation
- Vowels made use of the basic elements “earth” (horizontal line) and “humankind” (vertical line)

ㄱ “k” looks like the tongue root closing the throat

ㅜ “u” as in the middle sound of “jun”.

Design principles of Hangul



Summary

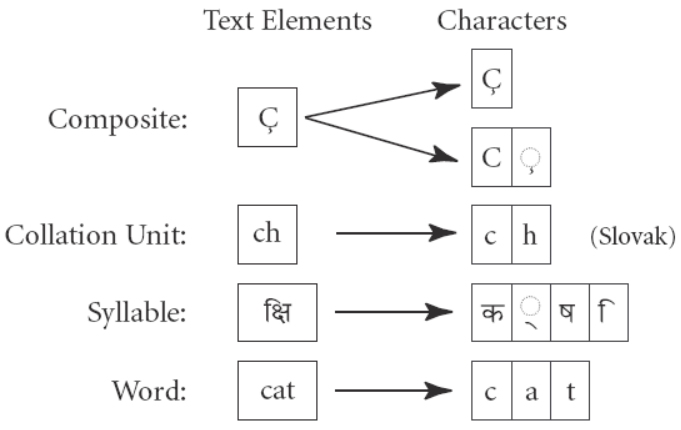
- Writing systems represent language in a variety of different ways
- But all writing systems represent sound to some degree
- While syllabaries are indeed common, virtually all syllabaries require some analysis below the syllable level

Encodings: Unicode

- Character encodings are arranged into "planes"
  - A plane consist of 65,536 (10000_{16}) "code points"
  - There are 17 planes (0-16) with Plane 0 being the "Basic Multilingual Plane"
- Texts are encoded in "logical" order, which is more abstract than the presentation order

Types of code points

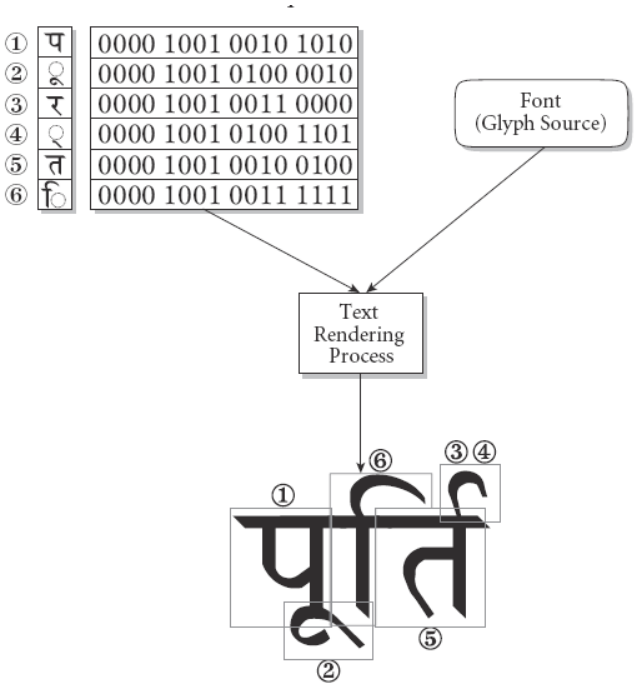
Basic Type	Brief Description	General Category	Character Status	Code Point Status
Graphic	Letter, mark, number, punctuation, symbol, and spaces	L, M, N, P, S, Zs	*Assigned to abstract character*	*Designated (assigned) code point*
Format	Invisible but affects neighboring characters; includes line/paragraph separators	Cf, Zl, Zp		
Control	Usage defined by protocols or standards outside the Unicode Standard	Cc		
Private-use	Usage defined by private agreement outside the Unicode Standard	Co		
Surrogate	Permanently reserved for UTF-16; restricted interchange	Cs	*Not assigned to abstract character*	
Noncharacter	Permanently reserved for internal usage; restricted interchange	Cn		
Reserved	Reserved for future assignment; restricted interchange			*Undesignated (unassigned) code point*



Example: Devanagari Code Points

Figure 2-2. Characters Versus Glyphs

Glyphs	Unicode Characters
A 𝐴 𝐴 𝐀 A 𝐴 𝐴 A	U+0041 LATIN CAPITAL LETTER A
a 𝑎 𝑎 **a** a 𝑎 𝑎 **a**	U+0061 LATIN SMALL LETTER A
fi fi	U+0066 LATIN SMALL LETTER F + U+0069 LATIN SMALL LETTER I
п *n* $\bar{u}$	U+043F CYRILLIC SMALL LETTER PE
ه ه ه ه	U+0647 ARABIC LETTER HEH



Example of Logical Ordering: Tamil /hoo/

Q: Doesn't the Tamil syllable ஹோ cause a problem in Unicode? This syllable can be encoded in two ways. In one case the characters are out of order, so doesn't this cause problems in text comparison and parsing?

A: The syllable *can* be represented in two ways:

A. ஹ ோ
0BB9 0BCB

B. ே ஹ ா
0BC7 0BB9 0BBE

However, **Line B** above is *incorrect*. The two correct possibilities are the following:

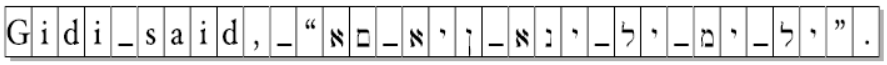
A. ஹ ோ
0BB9 0BCB

B'. ஹ ே ா
0BB9 0BC7 0BBE

UTF-8

- Common encoding of Unicode.
  - Variable length depending upon which code points one is dealing with
  - Programming languages have libraries that make dealing with UTF-8 strings easy.
  - Makes it easy to mix-and-match text from various sources:
    - , , , մայրաքաղաք,

Bidirectional text

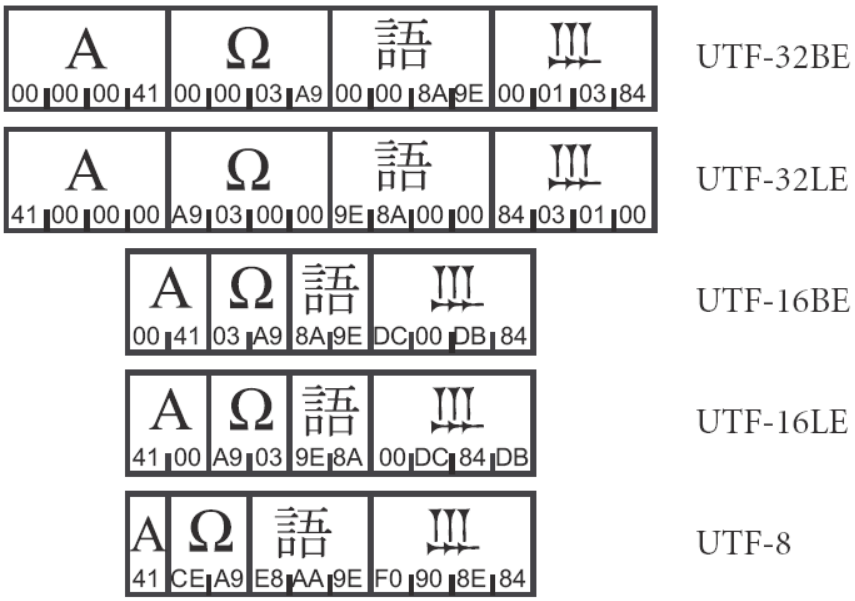


Unicode encoding schemes

Table 2-3. The Seven Unicode Encoding Schemes

Encoding Scheme	Endian Order	BOM Allowed?
UTF-8	N/A	yes
UTF-16 UTF-16BE UTF-16LE	Big-endian or Little-endian Big-endian Little-endian	yes no no
UTF-32 UTF-32BE UTF-32LE	Big-endian or Little-endian Big-endian Little-endian	yes no no

Figure 2-12. Unicode Encoding Schemes



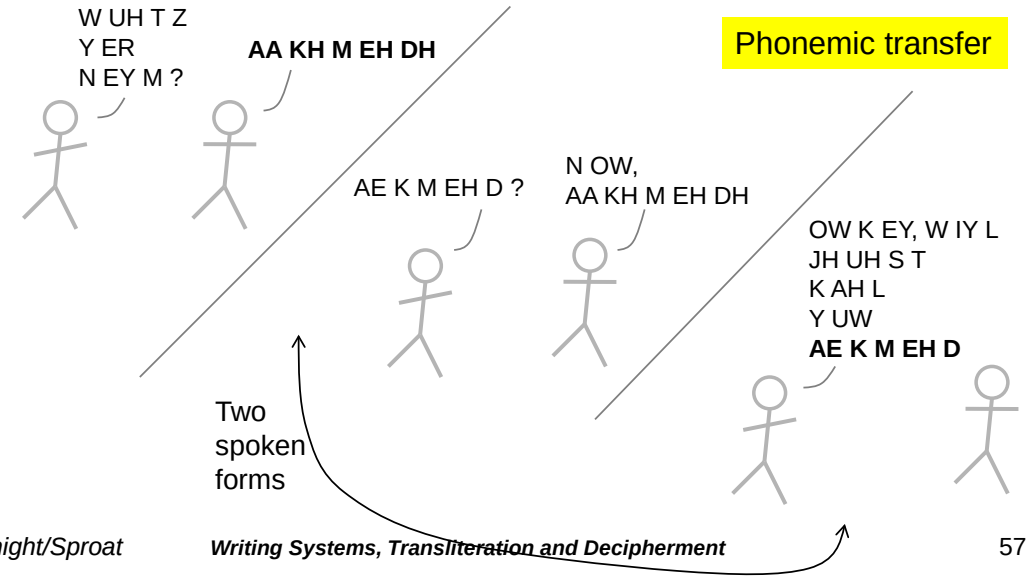
Issues with Unicode

- The design principles are nice, but they are inconsistently applied:
  - In Brahmi-derived alphasyllabaries each consonant and vowel has a separate code point.
  - Not so in Ethiopic
- In Indian alphasyllabaries, *logical order* is strictly enforced
  - Not so in Thai and Lao
- As we saw in the Tamil example, Unicode allows for variants for encoding the same information
- The term *ideograph* should never have become enshrined as the term for Chinese characters

Part II
Transcription (Transliteration)

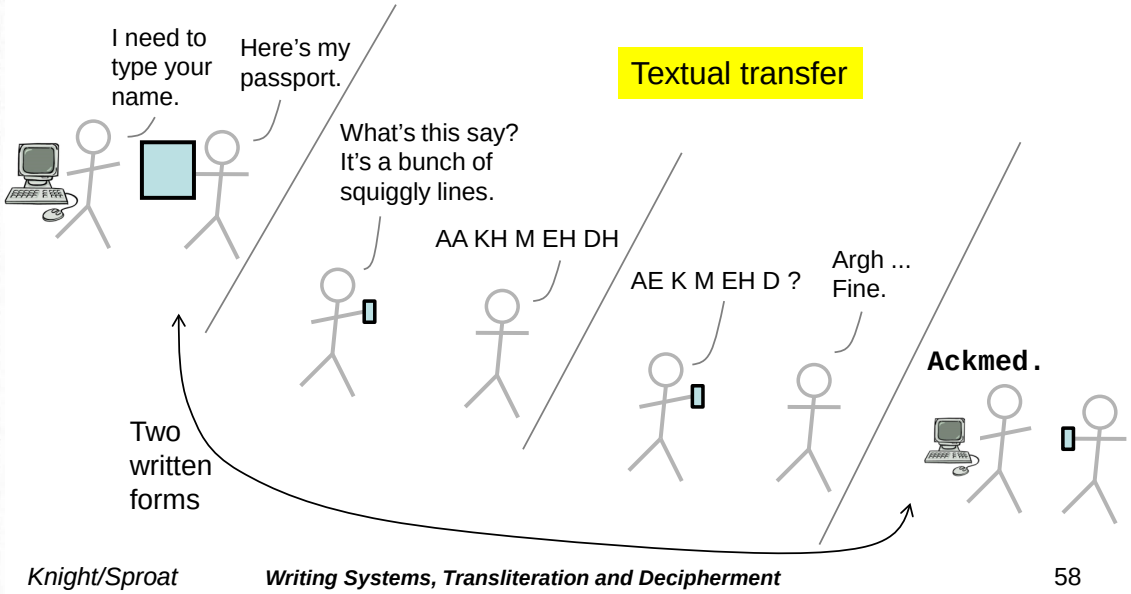
When Languages Collide

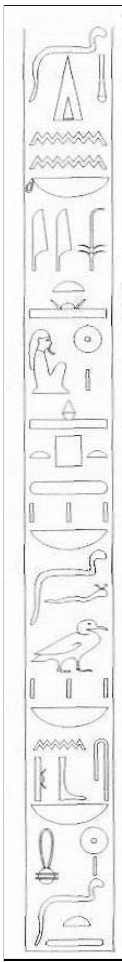
At the border crossing (before writing):



When Languages Collide

At the border crossing (<u>after</u> writing):



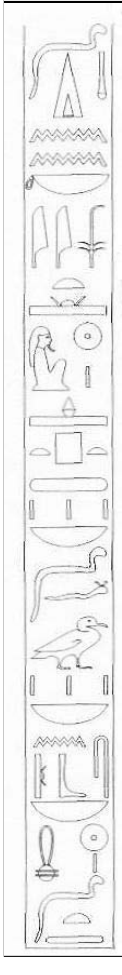


When Languages Collide

- Japanese/English example:

KEVIN	KNIGHT	English writing
K EH V IH N	N AY T	English sounds
K E B I N	N A I T O	Japanese sounds
ケビン	ナイト	Japanese writing

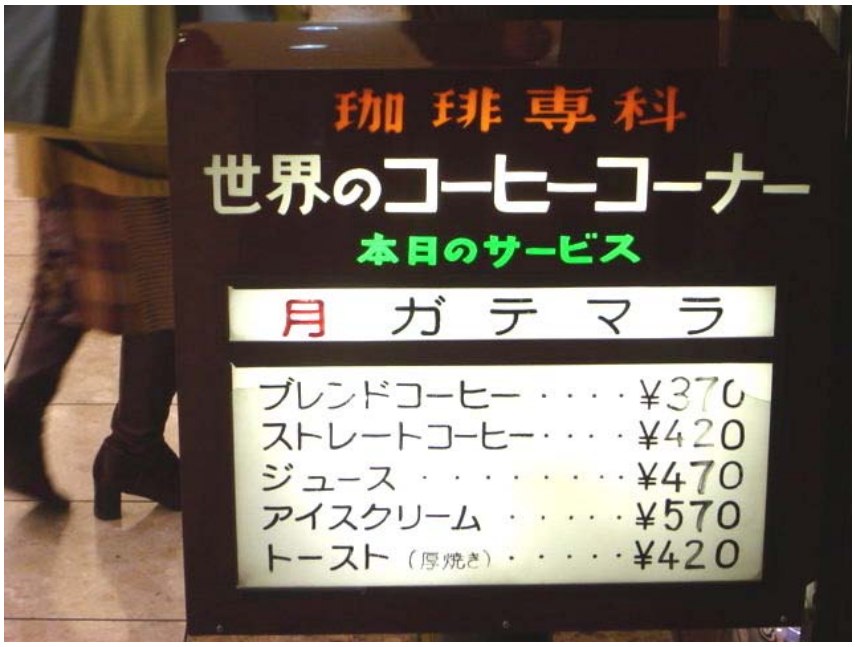
- V → B: phoneme inventory mismatch
- T → T O: phonotactic constraint
- alphabetic vs. syllabic writing



When Languages Collide

- **Common translation problem**
  - People and place names
  - New technical terms, borrowings
- **Challenging when source and target languages have:**
  - different phoneme inventories
  - different phonotactic constraints
  - different writing systems
- **English, Japanese, Russian, Chinese, Arabic, Greek …**

Streets of Tokyo / Katakana



Forward vs Backward Transcription

- Forward transcription
  - Import foreign term / name
    - Newt Gingrich → may be several ways to transcribe into Arabic
  - Generally flexible
- Backward transcription
  - Recover original term / name
  - Usually only one right answer
    - → Newt Gingrich (not Newt Kinkridge)

Japanese News

男子ゴルフ、米国ツアー・メジャー第1戦、マスターズ・トーナメント（2日目。1オーバーの51位タイからスタートした石川遼は、2バーディー、3ボギー、2ダブルボギーでスコアを5ストローク落とし、通算6オーバーの73位タイで予選落ちとなった

chyado kyanberu

首位には、7アンダーの単独首位からスタートし、5バーディー、3ボギーでスコアを2ストローク伸ばした米国のチャド・キャンベルと、4アンダーの6位タイからスタートし、5バーディー、ノーボギーでスコアを5ストローク伸ばした同じく米国のケニー・ペリーが通算9アンダーで並

kenii perii

2アンダーの21位タイからスタートした米国のタイガー・ウッズは、3バーディー、3ボギ

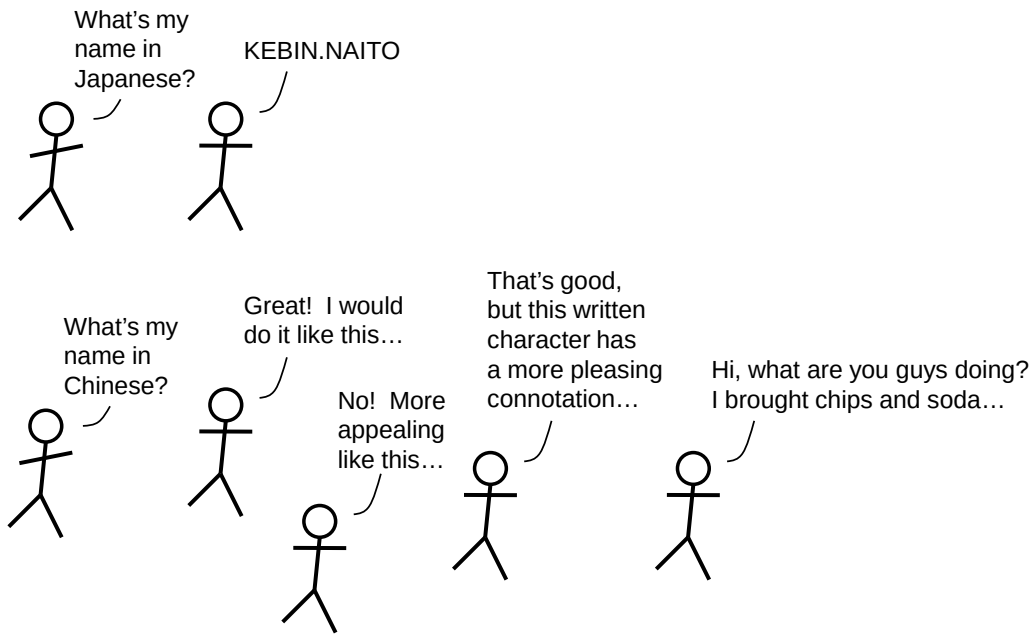
iibunpaa

ブンパーで2日目

taigaa uzzu

算2アンダーの18位タイにつけている。

Chinese/English



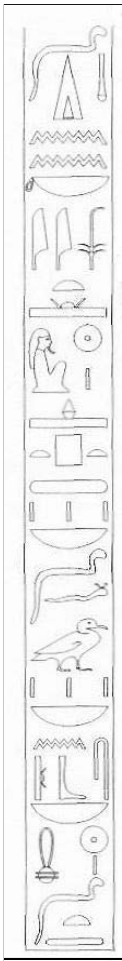
Chinese

- Several hundred syllables in inventory
  - Must stick to this idiosyncratic set
  - Washington → Hua Sheng Dun
  - No other syllables are easily written
- Homophony: after we decide on syllables, many characters to choose from
  - Washington → Hua Sheng Dun → 华 盛 顿
- Transcription vs Translation
  - Kevin Knight → Nai Kai Wen or Wu Kai Wen

Translation versus Transcription

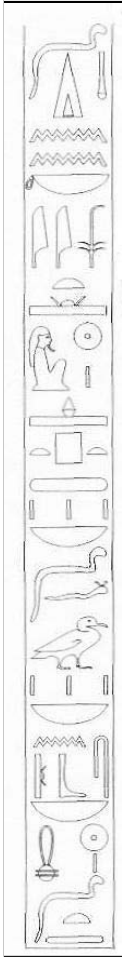
- Sometimes things are translated instead of transcribed
  - Japanese: computer → コンピューター (konpyuutaa)
  - Chinese: computer → **电脑** (dian nao) ("electric brain")
  - Arabic: Southern California → (Janoub Kalyfornya)

    ½ transliterated
    ½ translated



An Interesting Case: What's Going On Here?

- Observed English/Japanese transcription:
  - Tonya Harding → toonya haadingu
  - Tanya Harding → taanya haadingu
- Perhaps transcription is sensitive to source-language orthography …
- Or perhaps the transcriber is mentally mis-pronouncing the source-language word

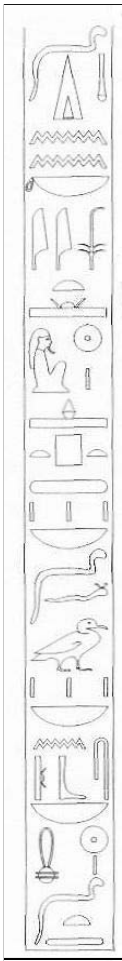


A Model of Transcription

KEVIN	KNIGHT	English writing
K EH V IH N	N AY T	English sounds
K E B I N	N A I T O	Japanese sounds
クヒ゛ン	ナイト	Japanese writing

Suppose we believe these are the steps.

We can model each step with a weighted finite-state transducer (WFST), and employ Claude Shannon's noisy-channel model.



A Model of Transcription

WFSA A

Angela Knight

WFST B

AE N J EH L UH N AY T

WFST C

a n j i r a n a i t o

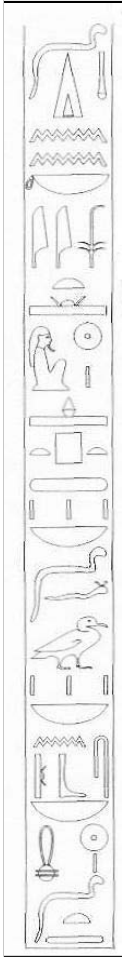
WFST D

アンジラナイト

MODELING DIRECTION

DECODING DIRECTION

[Knight & Graehl 98]



A Model of Transcription

WFSA A ← **LANGUAGE MODEL**

Angela Knight

WFST B ← **SPELLING TO SOUND TRANSDUCER**

AE N J EH L UH N AY T

WFST C ← **PHONEMIC TRANSFER TRANSDUCER**

a n j i r a n a i t o

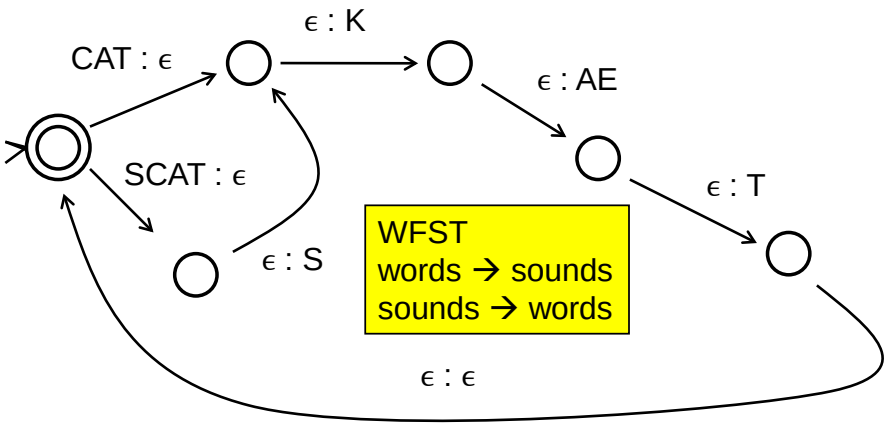
WFST D ← **SOUND TO SPELLING TRANSDUCER**

アンジラナイト

[Knight & Graehl 98]

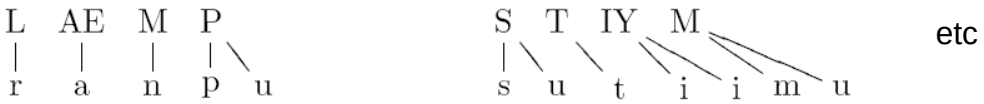
Spelling to Sound Transducer

- Richard talked about writing systems.
- Such a system captures an infinite relation of <sound-sequence, writing-sequence> pairs.



Learning Sequence Transformation Probabilities

Ideal training data:

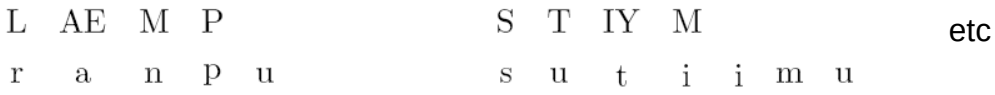


etc

P(n | M) = 0.5
P(m u | M) = 0.5

need much more data, of course

Actual training data:



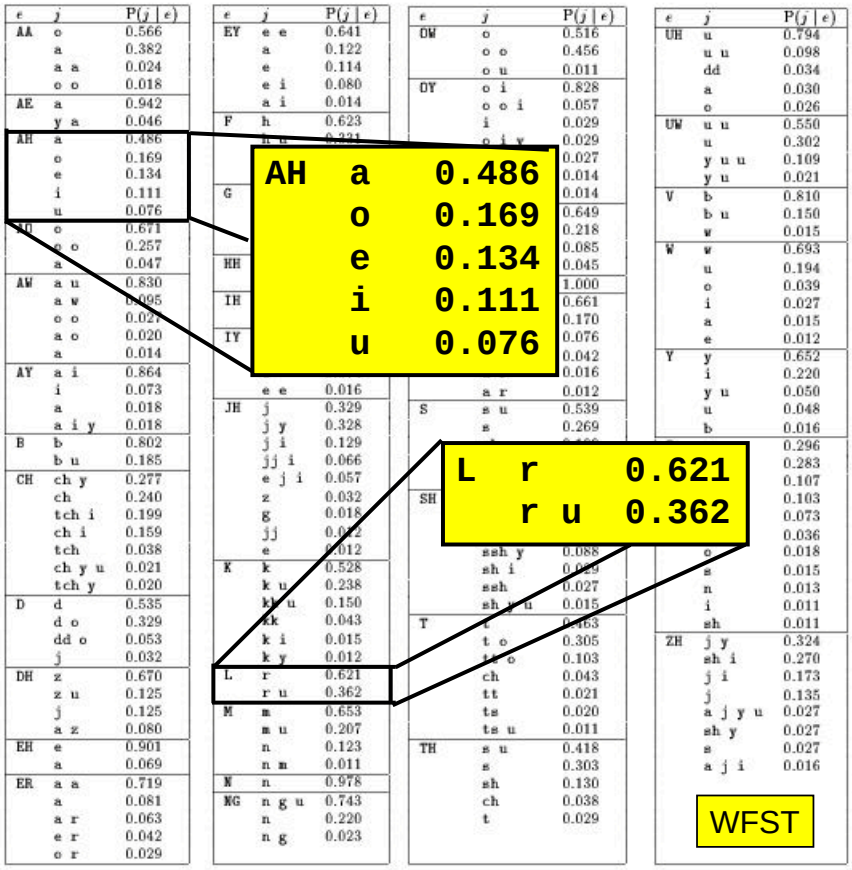
etc

Automatically align string pairs using the unsupervised Expectation-Maximization (EM) algorithm.

English-Japanese phonemic transfer patterns learned from parallel sequences

Learned by EM algorithm

[Knight & Graehl 98]



e	j	P(j | e)
AA	o	0.566
	a	0.382
	a a	0.024
	o o	0.018
AE	a	0.942
	y a	0.046
AH	a	0.486
	o	0.169
	e	0.134
	i	0.111
	u	0.076
AO	o	0.671
	o o	0.257
	a	0.047
AW	a u	0.830
	a w	0.095
	o o	0.027
	a o	0.020
	a	0.014
AY	a i	0.864
	i	0.073
	a	0.018
	a i y	0.018
B	b	0.802
	b u	0.185
CH	ch y	0.277
	ch	0.240
	tch i	0.199
	ch i	0.159
	tch	0.038
	ch y u	0.021
	tch y	0.020
D	d	0.535
	d o	0.329
	dd o	0.053
	j	0.032
DH	z	0.670
	z u	0.125
	j	0.125
	a z	0.080
EH	e	0.901
	a	0.069
ER	a a	0.719
	a	0.081
	a r	0.063
	e r	0.042
	o r	0.029

e	j	P(j | e)
EY	e e	0.641
	a	0.122
	e	0.114
	e i	0.080
	a i	0.014
F	h	0.623
	f u	0.231
G	g	0.986
	g u	0.014
HH	h	1.000
IH	i	0.649
	e	0.218
IY	i	0.085
	e e	0.045
	i y	0.042
	e	0.016
JH	j	0.329
	j y	0.328
	j i	0.129
	jj i	0.066
	e j i	0.057
	z	0.032
	g	0.018
	jj	0.012
	e	0.012
K	k	0.528
	k u	0.238
	kk u	0.150
	kk	0.043
	k i	0.015
	k y	0.012
L	r	0.621
	r u	0.362
M	m	0.653
	m u	0.207
	n	0.123
	n m	0.011
N	n	0.978
NG	n g u	0.743
	n	0.220
	n g	0.023

e	j	P(j | e)
OW	o	0.516
	o o	0.456
	o u	0.011
OY	o i	0.828
	o o i	0.057
	i	0.029
	o i y	0.029
	o y	0.027
	o	0.014
P	p	0.986
	p u	0.014
R	r	0.649
	r u	0.218
S	s	0.539
	s u	0.269
	a r	0.012
SH	sh	0.661
	sh u	0.170
	ssh y	0.088
	sh i	0.029
	ssh	0.027
	sh y u	0.015
T	t	0.653
	t o	0.305
	t o	0.103
	ch	0.043
	tt	0.021
	ts	0.020
	ts u	0.011
TH	s u	0.418
	s	0.303
	sh	0.130
	ch	0.038
	t	0.029

e	j	P(j | e)
UH	u	0.794
	u u	0.098
	dd	0.034
	a	0.030
	o	0.026
UW	u u	0.550
	u	0.302
	y u u	0.109
	y u	0.021
V	b	0.810
	b u	0.150
	w	0.015
W	w	0.693
	u	0.194
	o	0.039
	i	0.027
	a	0.015
	e	0.012
Y	y	0.652
	i	0.220
	y u	0.050
	u	0.048
	b	0.016
		0.296
		0.283
		0.107
		0.103
		0.073
		0.036
	o	0.018
	s	0.015
	n	0.013
	i	0.011
	sh	0.011
ZH	j y	0.324
	sh i	0.270
	j i	0.173
	j	0.135
	a j y u	0.027
	sh y	0.027
	s	0.027
	a j i	0.016

WFST

Angela Knight
Angela Nite
Andy Law Knight
Angela Nate *+ millions more*

WFSA A

Ann Gere Uh
Anne Jill Ahh
Angy Rugh
Ann Zillah *+ millions more*

WFST B

AE N J IH R UH N AY T
AH N J IH L UH N AY T OH
+ millions more

WFST C

a n j i r a n a i t o

WFST D

アンジラナイト

DECODING

A Model of Transcription

WFSA A

Angela Knight

WFST B

AE N J EH L UH N AY T

WFST C

a n j i r a n a i t o

WFST D

アンジラナイト

Can this transformation be learned from non-parallel data?

I.e., can katakana be deciphered without parallel text?

We'll return to this later → Decipherment section

Intermission

Alternative: Mapping Character Sequences Directly

KEVIN	KNIGHT	English writing
KE VI N	KN IGH T	English letter chunks
ケ ビ ン	ナ イ ト	Japanese writing

- Dispenses with spelling-to-sound models and pronunciation dictionaries
- Can be learned from parallel data using statistical MT-like techniques (over characters instead of words)

Hybrid Mapping Models

- Sound-based and character-based methods can be combined
  - [Al-Onaizan & Knight 02]
  - [Bilac & Tanaka 04, 05]
  - [Oh & Choi 2005, Oh et al 06]

Re-ranking Transcription Candidates

- Co-reference can help
  - Short name may be disambiguated by full version that appears earlier in a document
- Web counts can help
  - Bell Clinton (6m), Bill Clinton (27m)
- Context can help
  - Donald Martin » Donald Marron … but:
  - Donald Martin + Lightyear Capital (7)
  - Donald Marron + Lightyear Capital (6000)

[Al-Onaizan & Knight 02]

Use of Transcription in Machine Translation Systems

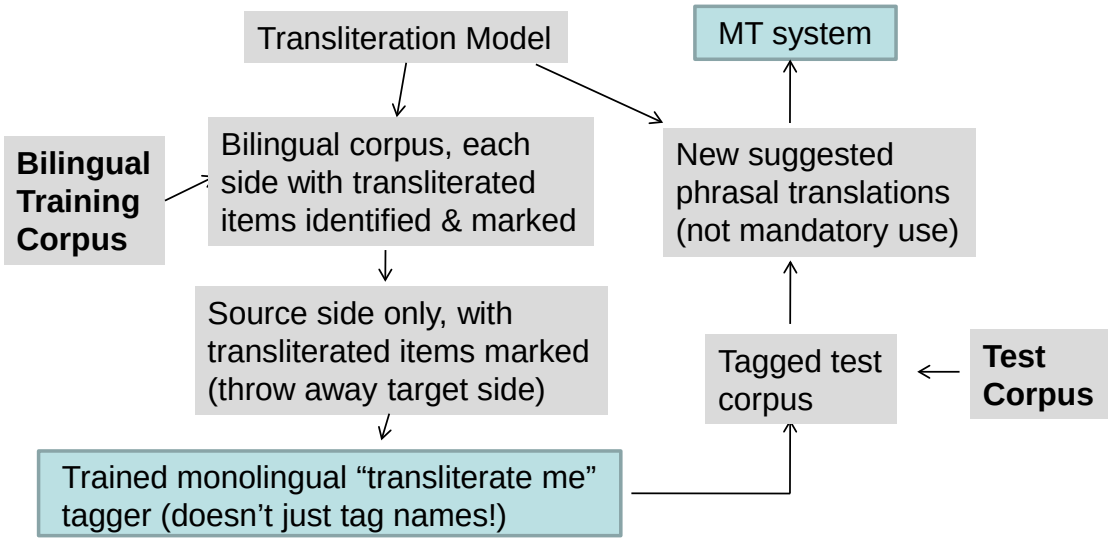
- What doesn't work:
  - Execute named-entity (NE) recognition on source text
  - Transcribe recognized items
  - Tell MT system to use transcriptions
- Often breaks a translation that was perfect before!
  - NE recognition is error-ful
  - Transcription is error-ful
  - Not all NEs should be transcribed
  - Phrase disruption
    - Vanilla MT system: … [f1 f2 f3] … → … [e1 e2 e3] …
    - "Improved" MT system: … f1 [f2 f3] … → … e5 [e2 e3] …

Whole phrase translation

NE ID + transcription

Use of Transcription in Machine Translation Systems

Another approach [Hermjakob et al 08]



Other Uses of Transcription Models

- Cross-lingual Information retrieval, eg, [Gao et al 04]
- Recognize transcriptions in comparable corpora, eg, [Sproat et al 06]
- Regional studies, eg, [Kuo et al 09]
- Automatic speech recognition
  - Phonemic transfer models might adjust for non-native speakers?
- Normalization of informal Internet Romanization schemes
  - Greek, Arabic, Russian
    - http://www.translatum.gr/converter/**greeklish-converter**.htm

Cypriot Greeklish with Instant Messaging Shorthand:	**`ego n 3ero re pe8kia.. skeftoume skeftoume omos tpt..`**
Normalized for automatic indexing or translation:	Εγώ εν ξέρω ρε παιθκιά... σκέφτουμαι σκέφτουμαι όμως τίποτα...

see "Greeklish", Wikipedia

Overview of the Transliteration/Transcription Literature

We have only touched on what is a large literature.

http://www.cs.mu.oz.au/~skarimi/

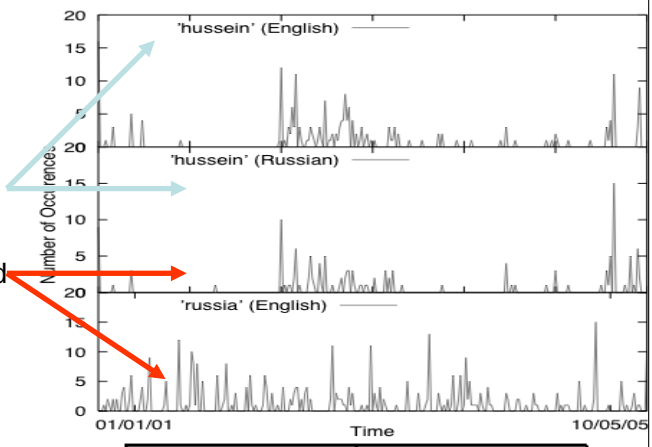
S. Karimi, F. Scholer, A. Turpin, A Survey on Machine Transliteration Literature, (Submitted Dec 08, Review received 31 Mar 09) Under Revision for ACM Computing Surveys.

Discriminative models

- Often used in judging potential transcription pairs in comparable corpora since here one is merely trying to classify the pair
- We will briefly review two pieces of work:
  - Klementiev & Roth 2006
  - Some results from the 2008 JHU summer workshop

Klementiev & Roth 2006

- Named entities (NEs) in one language co-occur with their counterparts in the other
  - *Hussein* has similar temporal histogram in both corpora
  - Different from histogram of word *Russia*
- NEs are often transcribed
- Approach is an iterative algorithm which exploits these two observations
- Given a bilingual corpus one side of which is tagged, it discovers NEs in the other language



English NE	Russian NE
lilic	лилич
fletcher	флетчер
bradford	брэдфорд
isabel	изабель
hoffmann	гофман
kathmandu	катманду

Klementiev & Roth 2006

- A *linear discriminative* approach for transcription model *M*
  - Use the perceptron algorithm to train *M*
  - The model activation provides the score used to select best transcriptions
  - Initialize *M* with a (small) set of transcriptions as positive examples and non-NEs paired with random words from *T* as negative examples

Klementiev & Roth 2006

- Features for the linear model *M* are:
  - For a pair of NE and a candidate (E_S, E_T) partition E_s and E_T into substrings of length 0 to n
  - Each feature is a pair of substrings
  - For example, (E_S, E_T) = (powell, pouel), n = 2
    - $E_s \rightarrow$ {_, p, o, w, e, l, l, po, ow, we, el, ll}
    - $E_T \rightarrow$ {_, p, o, u, e, l, po, ou, ue, el}
    - Feature vector is thus ((p,_), (p, a),… (w, au),… (el, el),…(ll, el))
- Use an observation that transcription tends to preserve phonetic sequence to limit the number of features
  - E.g. disallow couplings whose starting positions are too far apart (e.g. (p, ue) in the above example).

Klementiev & Roth 2006

Input

Bilingual comparable corpus (***S***,***T***)

Set of Named Entities in ***S***

Initialization

Initialize transcription model ***M***

Repeat

D ← ∅

For each NE in S

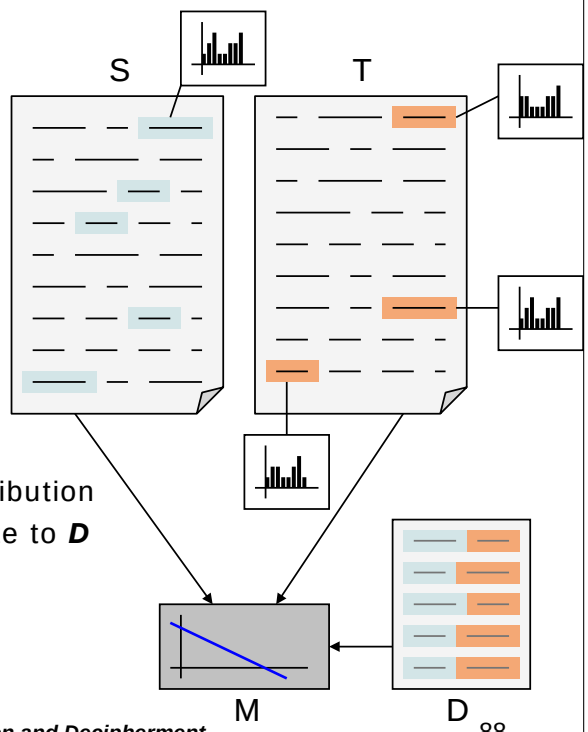
Collect candidates in ***T*** with high score (according to current ***M***)

For each candidate, collect time distribution

Add best temporally aligned candidate to ***D***

Use ***D*** to train ***M***

Until D stops changing



Klementiev & Roth

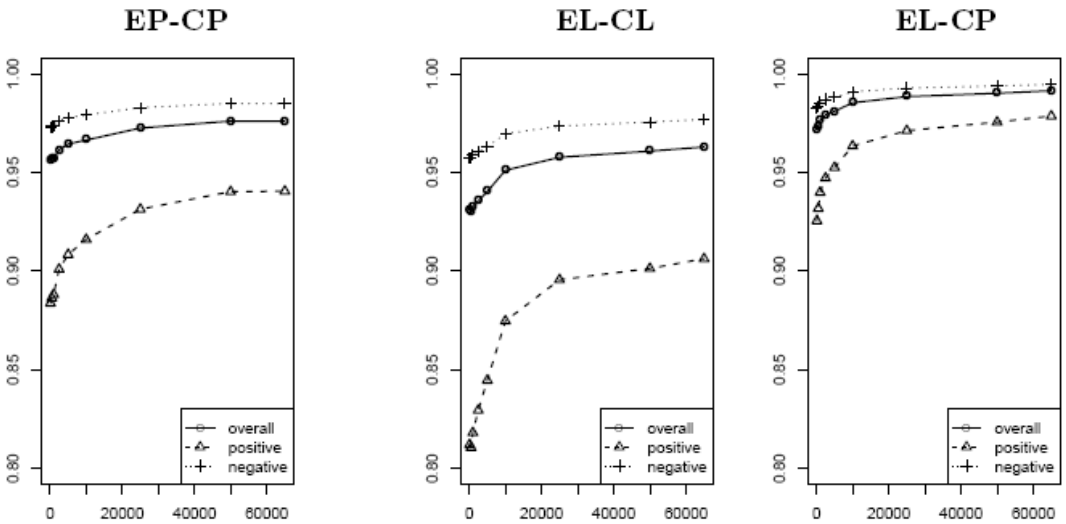
Algorithm iteratively refines transcription model with the help of time sequence similarity scoring

- Current transcription model chooses a list of candidates
- Best temporally aligned candidate is used for next round of training

	Iteration 0		Iteration 7
1	скорс {-е, -й, -йшего, -йший}	1	*форсайт {-а, -, -у}*
2	оформ {-лено, -лении, -ил, -ить}	2	оформ {-лено, -лении, -ил, -ить}
3	кокрэйн {-а, -}	3	проры {-вом, -ва, -ли, -тых, -вы, . . . }
4	флоре {-нс, -нц, -, -нции}	4	фросс
	•	5	фоссет {-т, -та, -ту, -а, -у}
	•		•
24	*форсайт {-а, -, -у}*		•
	•		

Example transcription candidate lists for NE ***forsyth*** for two iterations [correct is *форсайт*]

Representation matters (don't simply conclude that one should build a model based solely on orthography)



Pinyin is a relatively abstract "phonemic" representation that is not a particularly accurate representation of the pronunciation

Part III
Decipherment

Some decipherers

Thomas Young

Georg Friederich Grotefend

Jean François Champollion

Henry Creswicke Rawlinson

Michael Ventris

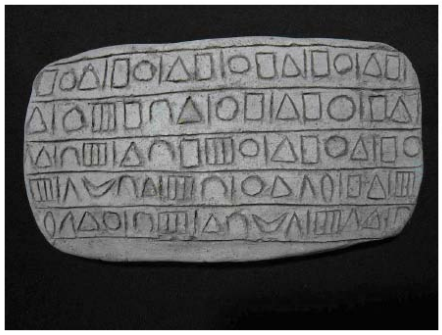
Not everything is decipherable

The Phaistos Disk:

Most serious scholars think the text is too short

A recent “find” from Jiroft (Iran)

Many suspect this Is a fake



Not everything that consists of linearly arranged symbols is writing

Symbols for the major deities of Aššurnaşirpal II

Not every communicative symbol system is writing

Naxi text

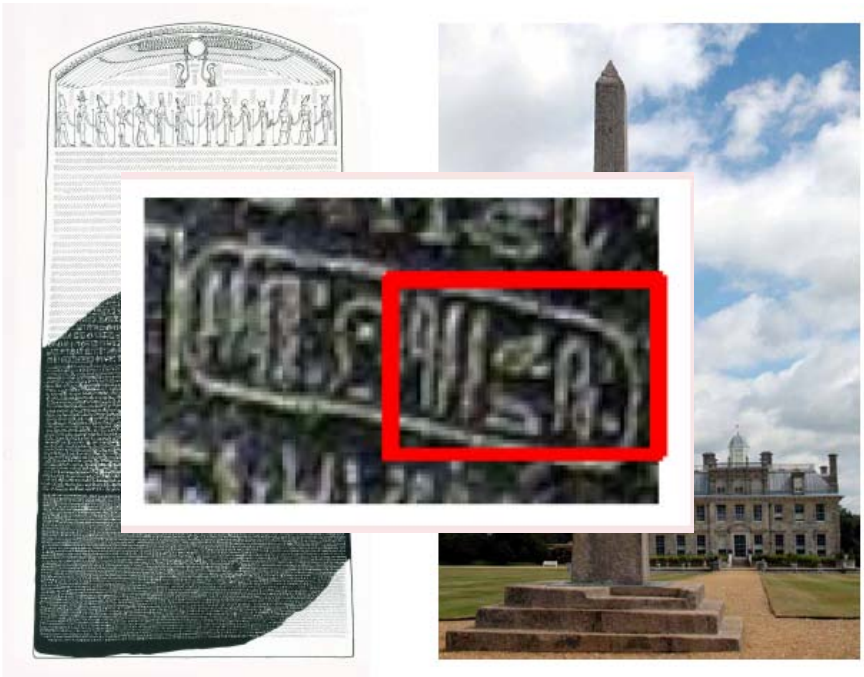
Questions that have to be asked

- Is the artifact genuine?
- Is the symbol system linguistic or non-linguistic?
  - If you have bilingual text that can help answer the question
- What is the underlying language?
- Which direction was the text read in?
- What kind of writing system are we dealing with?

Techniques and issues

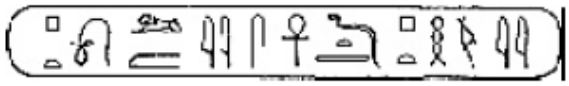
- Bilingual texts; names
- Structural analysis
- Verification

Parallel and comparable texts in Egyptian (Young & Champollion, 1816 onwards)



Parallel and comparable texts in Egyptian

p l yy s



t w m

Figure 4.12: *Ptolemy* from the Rosetta stone.

k p 3 t 3 t

l y w r

Figure 4.13: *Cleopatra* from the Bankes Obelisk.

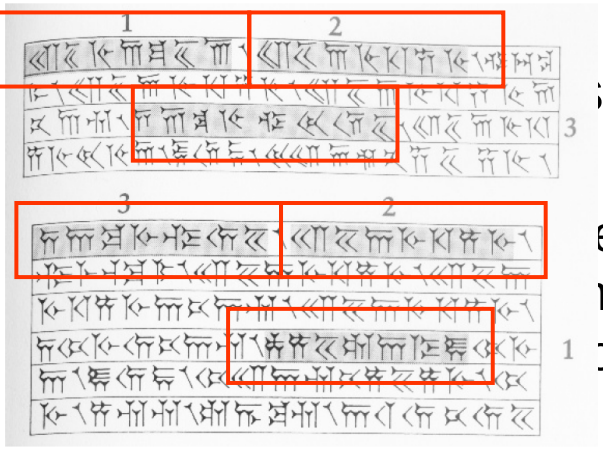
ptwlmyys (ΠΤΟΛΕΜΑΙΟΣ) djdt ‘nx mryy pth
Ptolemy, beloved of Ptah, may he be given (eternal) life

Parallel text --- without parallel text

- In early September 2008, many people were focussed on Hurricane Gustav, and what damage it might inflict upon the US oil industry in the Gulf of Mexico, or on the city of New Orleans…
- If you looked in Chinese newspapers at that time you’d find mention of **古斯塔夫** (*gǔsītǎfū)*
- Proper names are often an implicit source of parallel text

Grotefend's (1800) decipherment of Old Persian

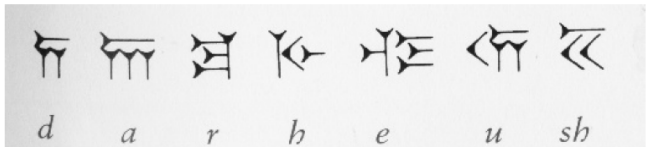
- Grotefend expected to find the names *Darius* and *Xerxes* in an inscription from Persepolis
- Grotefend g … in the inscription w …
  - By the large … en the separators … n must be alphabe … be wrong.)



Grotefend's decipherment of Old Persian

- From later Persian (Avestan) texts a few things were known:
  - Kings were designated in a very formulaic way: *X, great king, king of kings … son of Y*
  - *Xerxes* and *Darius's* names were something like ***xšherše*** and ***darheuš***
  - The later word for 'king' was ***kšeio***
- From history it was known that Xerxes was the son of Darius, and Darius the son of Hystapes (who was not a king)
- Grotefend reasoned the inscriptions might be:
  - *Xerxes great King … son of Darius*
  - *Darius great King … son of Hystapes*

A, great king, son of B
B, great king son of C



da-a-ra-ya-va-u-š / *xa-š-a-ya-θa-i-ya* / *va-za-ra-ka* / *xa-ša-a-ya-θa-i-ya* / *xa-ša-a-ya-θa-i-ya-a-na-a-ma* / *xa-ša-a-ya-θa-i-ya* / *da-ha-ya-u-na-a-ma* / ***vi-i-ša-ta-a-sa-pa***-*ha-ya-a* / *pa-u-ša* / *ha-xa-a-ma-na-i-ša-i-ya* / *ha-ya* / *i-ma-ma* / *ta-ca-ra-ma* / *a-ku-u-na-u-ša*

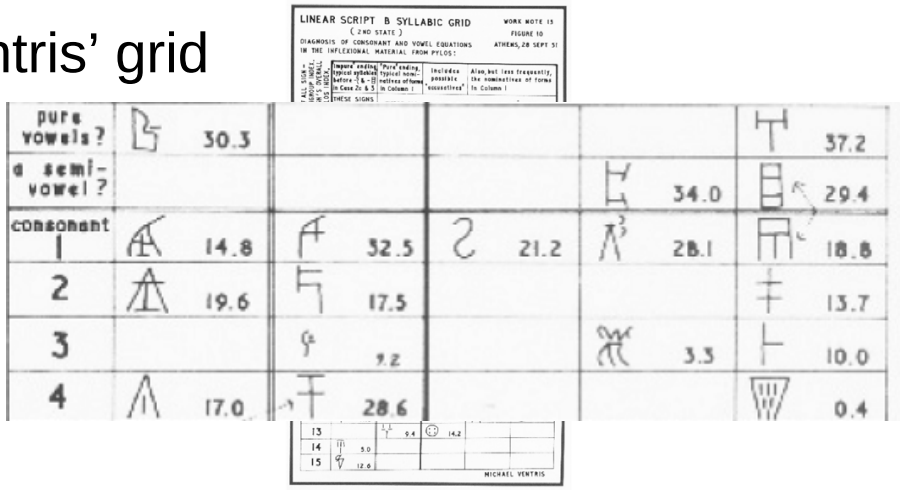
Darius, the great king, king of kings, king of countries, son of Hystapes, and Achaemenian, who built this palace.

Structural analysis: Linear B (Michael Ventris, early 1950's)

- Kober's "triplets":

ru ki to	a mi ni so
ru ki ti jo	a mi ni si jo
ru ki ti ja	a mi ni si ja
Luktos	*Amnisos*

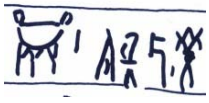
- Ventris' grid



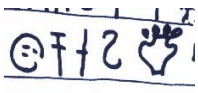
Verification: Linear B

- The phonology of many words corresponded to what was suspected for Greek from the relevant period:
  - *wa-na-ka (*wanaks,* later *anax* `ruler')
  - *i-qo* (**iqq*^{w}*os,* later *hippos* `horse')
- No definite articles
- Confirmation from new finds by Carl Blegen:

ti ri po de



qe to ro we



q^{w}etrōwes

↓

tetr-

Verification: Babylonian

- Babylonian is a complex mixed script.
  - The decipherment by Henry Creswicke Rawlinson and others seemed so arcane that many people doubted the decipherment
- In 1857 the Royal Asiatic Society received a letter from W.H. Fox Talbot containing a *sealed* translation of a text from the reign of Tiglath Pileser I (Middle Assyrian period, 1114–1076 BC)
- Talbot proposed comparing this with Rawlinson's translation, which was soon to be published
- Rawlinson not only agreed with this proposal, but suggested that two further scholars — Edward Hincks and Jules Oppert — be asked to provide translations.

Verification: Babylonian

Rawlinson: Then I went on to the country of Comukha, which was disobedient and witheld the tribute and offerings due to Ashur my lord.

Talbot: Then I advanced against Kummikhi, a land of the unbelievers who had refused to pay taxes and tribute unto Ashur, my lord.

Hincks: At that time I went to a disaffected part of Cummukh, which has withheld the tribute by weight and tale belonging to Assur, my lord.

Oppert: In these days I went to the people of Dummukh, the enemy who owed tribute and gifts to the god Asur, my lord.

Verification
A text from Abu Simbel

ra ms S S

Tuthmosis

On the Rosetta Stone, (2) was found to be aligned with the Greek word *genethlia* 'birthday': the Coptic word for birth was *mīse* confirming the *ms* reading for this glyph

How complete must a decipherment be for it to be verified?

153 = Er02 (with addition of new fragments) [880]

1 [*e*]-*ke*-*ra*$_2$-[*wo* *ki*]-*ti*-*me*-*no* *e*-*ke*
2 *sa*-*ra*-*pe*-*do*-[*i* ? *pe*]-*pu*$_2$?-*te*-*me*-*no*
3 *to*-*so* [*pe*-*ma*] WHEAT 30[+20?]
4 *to*-*so*-*de* [...]-*to* *pe*-*ma* WHEAT 42[+2?]
5 *to*-*sa* *we*-*je*-[*we*] 1100[
6 *to*-*sa*-*de* *ṣu*-*za* [?] 1000[
vacat
8 *ku*-*su*-*to*-*ro*-*pa*$_2$ *to*-*ṣo* *pe*-*ma* 94

?Ekhelāwōn has *private* (lands) on the S~peda, planted with trees.
So much seed: ?6000 l. wheat,
so much seed of the [...]: ?5280 l. wheat.
So many [...]: 1100?
So many fig-trees: 1000?
Aggregate, so much seed: 11,280 l.

From Ventris & Chadwick, *Documents in Mycenaean Greek*

Prospects for Automatic Decipherment

- Automatic decipherment is why computers were invented, in the 1940s
- Of course, military ciphers are different from unknown scripts
- But similar skills and techniques may apply

Letter Substitution Cipher

- Plaintext: **HELLO WORLD ...**
- Secret encipherment key:

```
PLAIN:   ABCDEFGHIJKLMNOPQRSTUVWXYZ
CIPHER:  PLOKMIJNUHBYGVTFCRDXESZAQW
```

- Ciphertext: **NMYYT ZTRYK ...**

- Key is unknown to code-breaker
- What key, if applied to the ciphertext, would yield sensible plaintext?

. . . .

KDCY LQZKTLJQX CY MDBCYJQL: “TR

. . . . . .

HYD FKXC, FQ MKX RLQQIQ HYDL

. . . .

MKL DXCTW RDCDLQ JQMNKXTMB

. . . . .

PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######### V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

a o e.a .e o o.e .

KDCY LQZKTLJQX CY MDBCYJQL: “TR

.o .a .e a . ee.e .o

HYD FKXC, FQ MKX RLQQIQ HYDL

a . . e .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.e a .a. o e.a

PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######### V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 6  #### V
Z 2  .
```

auto repairmen to customer if

KDCY LQZKTLJQX CY MDBCYJQL: “TR

you wait we can freeze your

HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics

MKL DXCTW RDCDLQ JQMNKXTMB

discover a way to repair it

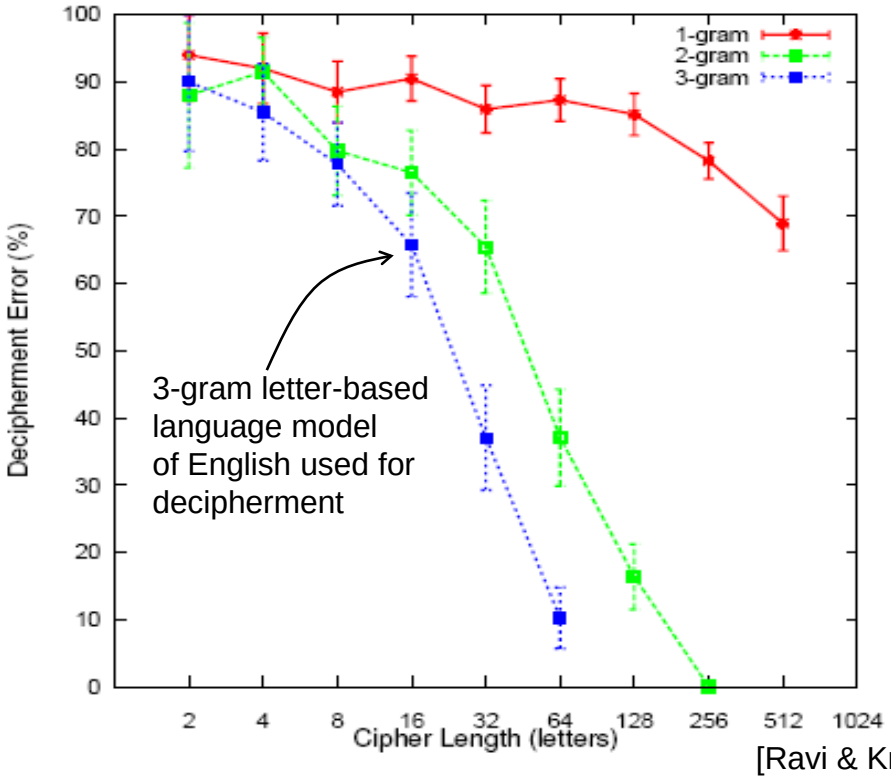
PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######### V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 6  #### V
Z 2  .
```

Letter Substitution Cipher

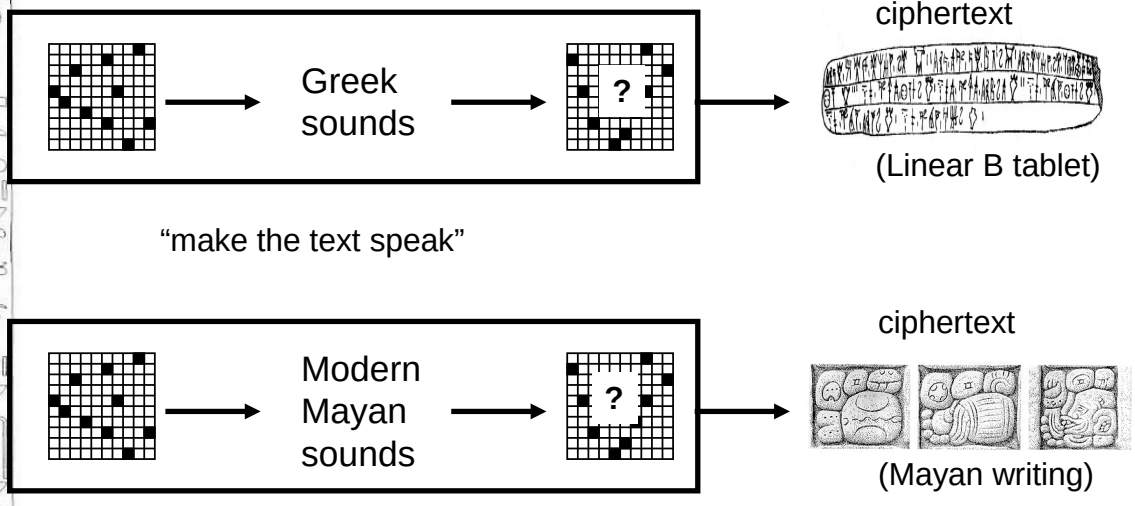
- How little knowledge of the plaintext language is necessary for decipherment?
  - Simple letter-based n-gram models
  - P(a | t) -- given t, chance that next letter is a
- EM-based decipherment
  - [Knight et al 06]
- Integer-programming-based decipherment
  - [Ravi & Knight 08]

Letter Substitution Cipher

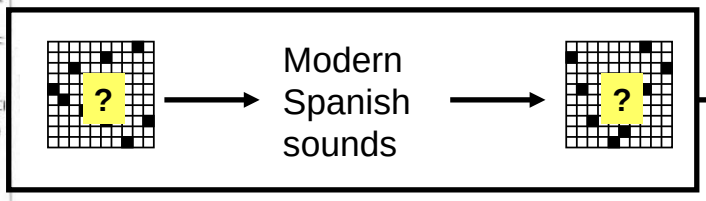


[Ravi & Knight 08]

Unknown Script as a Cipher



Unknown Script as a Cipher



ciphertext (6980 letters)

primera parte
del ingenioso
hidalgo don …

(Don Quixote)

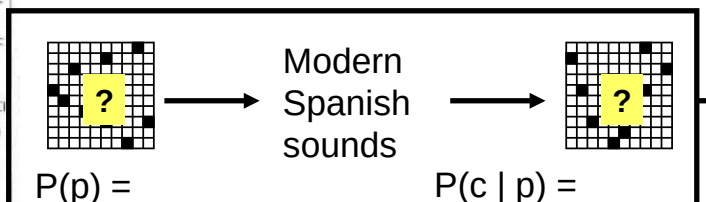
26 sounds:
B, D, G, J (ca<u>ny</u>on),
L (<u>y</u>arn), T (<u>th</u>in), a,
b, d, e, f, g, i, k, l,
m, n, o, p , r,
rr (trilled), s,
t, tS, u, x (h<u>a</u>t)

?

32 letters:
ñ, á, é, í, ó, ú,
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

[Knight & Yamada, 1999]

Unknown Script as a Cipher



P(p) =
P(p1 | START) *
P(p2 | p1) *
P(p3 | p2) * …

Phoneme
bigram model
P(L | tS) = 0.003

P(c | p) =
P(c1 | p1) *
P(c2 | p2) *
P(c3 | p3) * …

Phoneme-to-
letter model
P(y | L) = 0.8 ?

ciphertext (6980 letters)

primera parte
del ingenioso
hidalgo don …

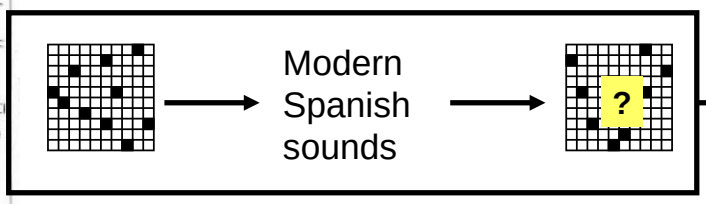
(Don Quixote)

Ideal “Key”

sound	letter
B	b or v
D	d
G	g
J	ñ
L	l l or y
a	a or á
b	b or v
d	d
e	e or é
f	f
g	g
i	i or í
l	l
m	m
n	n
o	o or ó
p	p

sound	letter
r	r
t	t
tS	c h
u	u or ú
x	j
nothing	h
T (before a, o, u)	z
T (before e or I)	c or z
T (otherwise)	c
k (before e or I)	q u
k (before s)	x
k (otherwise)	c
rr (start of word)	r
rr (otherwise)	rr
s (after k)	nothing
s (otherwise)	s

Unknown Script as a Cipher



ciphertext (6980 letters)

primera parte del ingenioso hidalgo don …

(Don Quixote)

EM-based decipherment finds a very good “key” and achieves 93% phoneme accuracy

Correct sounds: **primera parte del inxenioso iDalGo don kixote…**
Deciphered sounds: **primera parte del inGenioso biDalGo don kixote…**

How to Decipher Unknown Script if Spoken Language is Also Unknown?

- One idea: build a *universal* model P(s) of human phoneme sequence production

- Human might generally say: K AH N AH R IY
- Human won't generally say: R T R K L K

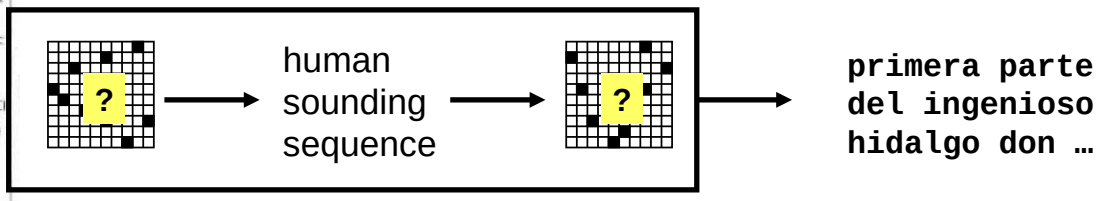
- Deciphering means finding a P(c | p) table such that there is a decoding with a good universal P(p) score

Universal Phonology

- Linguists know lots of stuff!
- Phoneme inventory
  - if z, then s
- Syllable inventory
  - all languages have CV (consonant-vowel) syllables
  - if VCC, then also VC
- Syllable sonority structure
  - {stdbptk} {mnrl} {V} {mnrl} {stdbptk}
  - dram, lomp, tra, ma, ? rdam, ? lopm, ? tba, ? mla
- Physiological preference constraints
  - tomp, tont, tongk, ? tomk, ? tonk, ? tongt, ? tonp

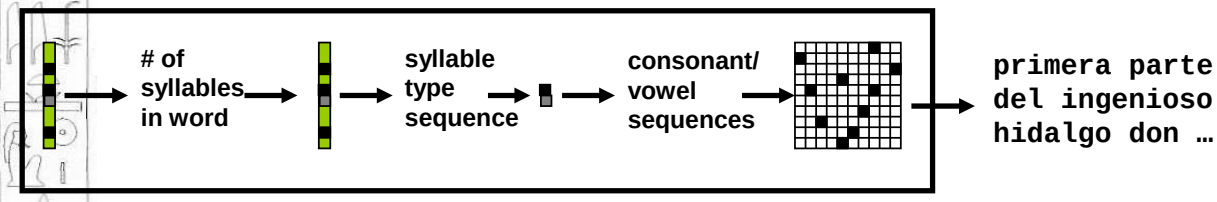
Universal Phonology

Task 1: Label each letter with a phoneme



Universal Phonology

Task 2: Label each letter with a phoneme class: C or V



P(1) = ?
P(2) = ?
etc.

P(CV) = ?
P(V) = ?
P(CVC) = ?
+ 7 other types

P(V | V) = ?
P(VV | V) = ?

P(a | V) = ?
P(a | C) = ?
etc.

Input: **primera parte del ingenioso hidalgo don …**
Output: **CCVCVCV CVCCV CVC VCCVCVVCV CVCVCCV CVC …**

P(CV)	= 0.45	P(VC)	= 0.09
P(V)	= 0.15	P(CVC)	= 0.22
P(CCV)	= 0.02	P(CCVC)	= 0.01

P(a | V)	= 0.27	P(a | C)	= 0.00
P(b | V)	= 0.00	P(b | C)	= 0.04
P(c | V)	= 0.00	P(c | C)	= 0.07

Unknown Source Language

- Another idea: brute force
- If we don't know the spoken language, simply decode against all spoken languages:
  - Pre-collect P(p) for 300 languages
  - Train a P(c | p) using each P(p) in turn
  - See which decoding run assigns highest P(c)
- Hard to get phoneme sequences
- Can use text sequence as a substitute

UN Declaration of Human Rights

Exists in many of world's languages, UTF-8 encoding

No one shall be arbitrarily deprived of his property
Niemand se eiendom sal arbitrêr afgeneem word nie
Asnjeri nuk duhet të privohet arbitrarisht nga pasuria e tij

Janiw khitisa utaps oraqeps inaki aparkaspati
Arrazoirik gabe ez zaio inori bere jabegoa kenduko
Den ebet ne vo tennet e berc'hentiezh digantañ diouzh c'hoant
Никой не трябва да бъде произволно лишен от своята собственост
Ningú no serà privat arbitràriament de la seva propietat

Di a so prupiità ùn ni pò essa privu nimu di modu tirannicu
Nitko ne smije samovoljno biti lišen svoje imovine
Nikdo nesmí být svévolně zbaven svého majetku
Ingen må vilkårligt berøves sin ejendom
Niemand mag willekeurig van zijn eigendom worden beroofd

Nul ne peut être arbitrairement privé de sa propriété
Nimmen mei samar fan syn eigendom berôve wurde
Ninguín será privado arbitrariamente da súa propiedade
Niemand darf willkürlich seines Eigentums beraubt werden
Κανείς δεν μπορεί να στερηθεί αυθαίρετα την ιδιοκτησία του
Avavégui ndojepe'a va'erâi oimeháicha reinte imbáe teéva
Ba wanda za a kwace wa dukiyarsa ba tare da cikakken dalili ba
Senkit sem lehet tulajdonától önkényesen megfosztani
Engan má eftir geðþótta svipta eign sinni
Necuno essera private arbitrarimente de su proprietate
Ní féidir a mhaoin a bhaint go forlámhach de dhuine ar bith
Al neniu estu arbitre forprenita lia proprieto
Kelleltki ei tohi tema vara meelevaldselt ära võtta
Eingin skal hissini vera fyri ongartøku
Me kua ni dua e kovei vua na nona iyau
Keltään älköön mielivaltaisesti riistettäkö hänen omaisuuttaan

Unknown Source Language

- Input:

  `cevzren cnegr qry vatravbfb uvqnytb qba dhvwbgr qr yn znapun …`

- Top 5 languages with best P(c) after deciphering:

-5.29120	spanish
-5.43346	galician
-5.44087	portuguese
-5.48023	kurdish
-5.49751	romanian

- Best-path decoding assuming plaintext is Spanish:

  `primera parte del ingenioso hidalgo don quijote de la mancha …`

- Best-path decoding assuming plaintext is English:

  `wizaris asive bek u-gedundl pubscon bly whualve be ks asequs …`

- Simultaneous language ID and decipherment

Transliteration as a Cipher

- Ciphertext: Japanese Katakana
- Plaintext: English

[Ravi & Knight 09]

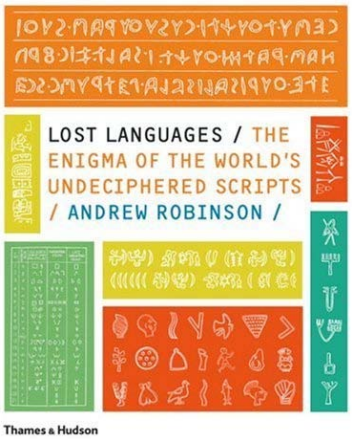
Foreign Language as a Cipher?

- Ciphertext: Billions of words of Albanian
- Plaintext: English

Is it possible to train statistical MT systems with little or no parallel text?

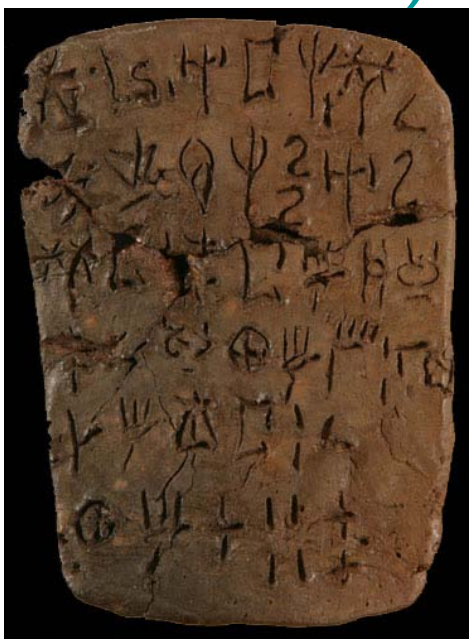
What's left to decipher?

- Proto-Elamite
- Linear A
- Etruscan
- *rongorongo*
- *Indus Valley*
- *Phaistos disk*
- Epi-Olmec and other Mesoamerican scripts



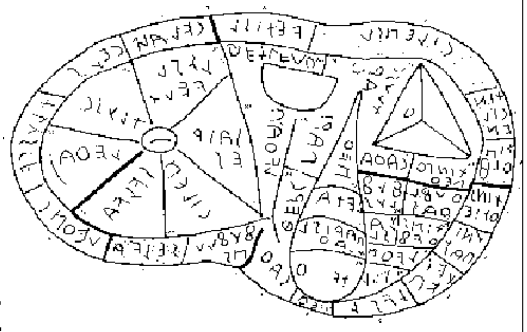
Linear A
(Crete, ca 2000 BC to 1200 BC)

- Clearly the precursor of Linear B
- Mostly accounting texts (like Linear B), though there are other kinds of inscriptions
- We can “read” the texts but we don’t know much about the underlying language.



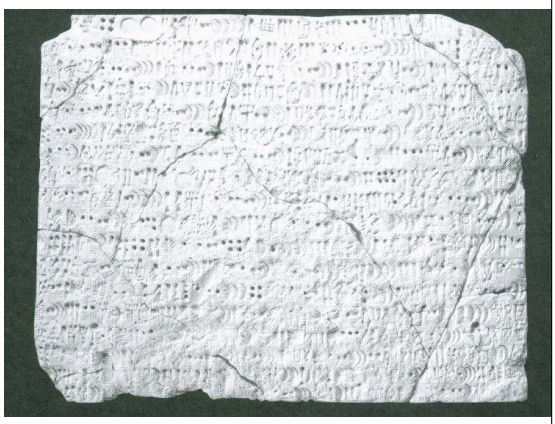
Etruscan
(Italy, 700 BC – 1st Century AD)

- The alphabet is known – it was derived from Greek and was the precursor to Latin
- The language (like that of Linear A) is largely unknown



Proto-Elamite (Iran, ca. 3100 – 2900 BC)

- Possibly as many as 5,500 distinct signs (?)
- Underlying language is unknown – may be Elamite (cf later linear Elamite inscriptions) but that is not clear



rongorongo (Easter Island – 19th Century)

- About 600 zoomorphic and anthropomorphic glyphs
- Extant corpus is about 12,000 glyphs long, all carved on driftwood
- The underlying language (Rapanui) is known
- Ethnographic accounts of the *rongorongo* ceremonies exist
- Claims to the contrary aside, there is *no evidence* this was a writing system in the normal sense.
  - The only bit of text that has been "deciphered" is a calendar

Indus Valley (South Asia, 26th—20th century BC)

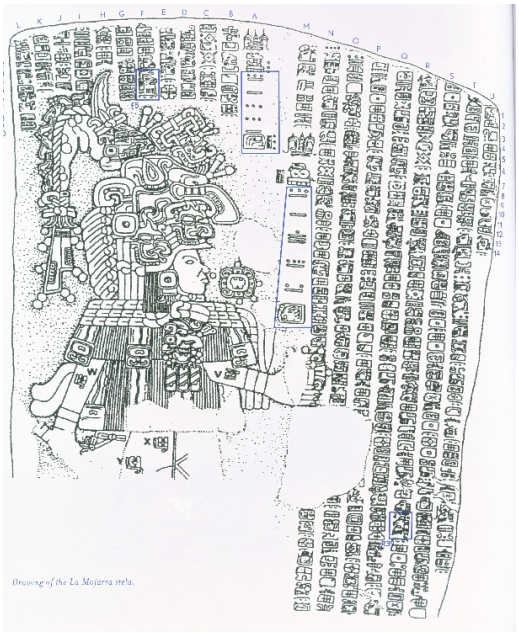
- System with a few hundred glyphs
- Inscriptions are *very* short – longest on a single surface has 17 glyphs
- The "standard" theory, due to Asko Parpola, is that this was a Dravidian language
- Recently, Farmer, Witzel and Sproat argued that this was not a writing-system at all

Phaistos disk (Crete, ca 1800 BC??)

- 241 tokens with 45 distinct glyphs
  - Glyphs are all pictographic – images of animals, people, various objects
- Text is on both sides of disk in a spiral working from the outside
- *The Phaistos Disk is the world's first known printed document*
- There has been a recent suggestion (by ancient art dealer Jerome Eisenberg) that it may be a fake
- In any case, the text is *too short* to allow for a verifiable decipherment

Epi-Olmec/Isthmian (Mesoamerica – 1400 BC??)

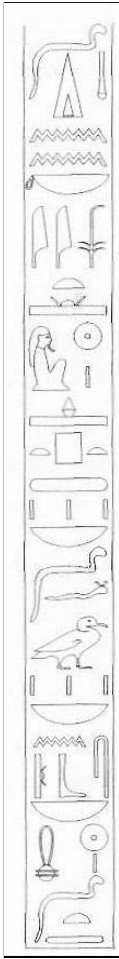
- About 600 characters of text extant
- Approx. 166 non-numerical signs
- Justeson and Kaufman proposed a decipherment as (epi)-Olmec in 1992
- But this is hotly contested …



Further Reading

Writing systems

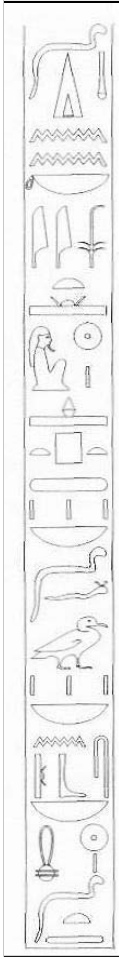
- P. Daniels, and W. Bright (editors). 1996. *The World's Writing Systems.* Oxford: Oxford University Press.
- H. Rogers. *Writing Systems: A Linguistic Approach*, Blackwell, 2005.
- A. Robinson. 2006. *The Story of Writing: Alphabets, Hieroglyphs and Pictograms*. Thames and Hudson, London.
- A. Gnanadesikan. 2008. *The Writing Revolution: Cuneiform to the Internet*. Wiley-Blackwell, Malden, MA.
- R. Sproat. *Language, Technology and Society*. Oxford, Oxford University Press, Forthcoming, 2009.



Further Reading

Encoding: there are many documents on the web that discuss encoding issues, including various documents from the Unicode Consortium.

However, one of the best starting places is:
http://www.joelonsoftware.com/articles/Unicode.html

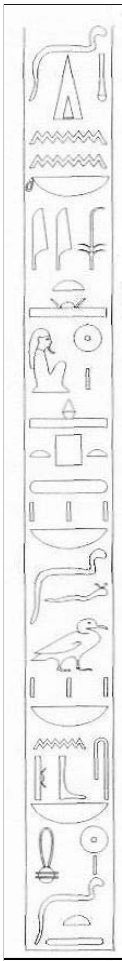


Further Reading

Transliteration/Transcription

http://www.cs.mu.oz.au/~skarimi/

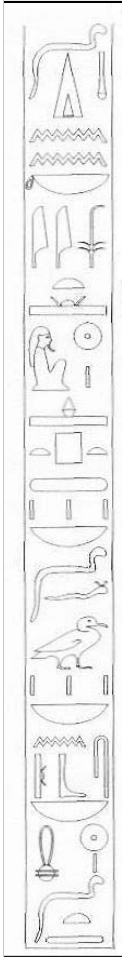
S. Karimi, F. Scholer, A. Turpin, A Survey on Machine Transliteration Literature, (Submitted Dec 08, Review received 31 Mar 09) Under Revision for ACM Computing Surveys.



Further Reading

Discriminative models of transcription

1. A. Klementiev and D. Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *ACL*.
2. D. Zelenko and C. Aone. 2006. Discriminative methods for transliteration. In *EMNLP.*
3. S-Y. Yoon, K-Y. Kim, and R. Sproat. 2007. Multilingual transliteration using feature based phonetic method. In *ACL*.
4. D. Goldwasser and D. Roth. 2008. Active sample selection for named entity transliteration. In *ACL.*



Further Reading

Decipherment

1. R. Parkinson. 1999. *Cracking Codes: The Rosetta Stone and Decipherment*. University of California Press, Berkeley.
2. M. Pope. 1999. *The Story of Decipherment: From Egyptian Hieroglyphs to Maya Script.* Thames and Hudson, London.
3. A. Robinson. 2002. *The Man who Deciphered Linear B: The Story of Michael Ventris*. Thames and Hudson, London.
4. A. Robinson. 2009. *Lost Languages: The Enigma of the World's Undeciphered Scripts.* Thames and Hudson, London.

Further Reading

Auto Decipherment

1. “A Computational Approach to Deciphering Unknown Scripts”, (K. Knight and K. Yamada), Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, 1999.
2. “Unsupervised Analysis for Decipherment Problems”", (K. Knight, A. Nair, N. Rathod, and K. Yamada), Proc. ACL-COLING (poster), 2006.
3. “Attacking Decipherment Problems Optimally with Low-Order N-gram Models”, (S. Ravi and K. Knight), Proc. EMNLP, 2008.
4. “Learning Phoneme Mappings for Transliteration without Parallel Data”, (S. Ravi and K. Knight), Proc. NAACL, 2009.

the end