Overview

• An overview of writing systems

• Transcription/transliteration between scripts

• Traditional and automatic approaches to decipherment
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”

\[ \text{mat d\text{og} l\text{atks} \text{æv\text{ə}k\text{a}d\text{oz}}} \]

\[ (\text{mat}) \ (\text{d\text{og}}) \ (\text{l\text{atks}}) \ (\text{æ})(\text{v\text{ə}})(\text{k\text{a}})(\text{d\text{oz}}) \]

\[ [\text{(æ)(v\text{ə})}][\text{(k\text{a})}(\text{d\text{oz}})] \]

\[ \text{mat d\text{og} l\text{atk}\text{+s} \text{æv\text{ə}k\text{a}d\text{oz}}} \]

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’
• **zhīshì** indicators
  – ‘above’, ‘below’
• **huìyì** meaning compound
  – ‘bright’ (SUN+MOON)
• **xīngshēng** 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’)

[79x31]Knight/Sproat
[79x31]Writing Systems, Transliteration and Decipherment
Xíngshēng characters

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

<table>
<thead>
<tr>
<th>Character</th>
<th>Pinyin</th>
</tr>
</thead>
<tbody>
<tr>
<td>鱼</td>
<td>yú</td>
</tr>
<tr>
<td>里 (‘fish’) = FISH · LI</td>
<td></td>
</tr>
<tr>
<td>鸟</td>
<td>niǎo</td>
</tr>
<tr>
<td>甲 (‘duck’) = BIRD · JIA</td>
<td></td>
</tr>
<tr>
<td>草</td>
<td>cǎo</td>
</tr>
<tr>
<td>早 (‘grass’) = VEGETATION · ZAO</td>
<td></td>
</tr>
<tr>
<td>志</td>
<td>zhì</td>
</tr>
<tr>
<td>心</td>
<td>xīn</td>
</tr>
<tr>
<td>士 (‘will, goal’) = HEART · SHI</td>
<td></td>
</tr>
<tr>
<td>国</td>
<td>guó</td>
</tr>
<tr>
<td>圈</td>
<td>quān</td>
</tr>
<tr>
<td>或</td>
<td>huó</td>
</tr>
<tr>
<td>甸 (‘country’) = ENCLOSURE · HUO</td>
<td></td>
</tr>
</tbody>
</table>

General operation: **SEMANTIC · PHONETIC**
General operation: **SEMANTIC - PHONETIC**

<table>
<thead>
<tr>
<th>Character</th>
<th>Meaning</th>
<th>Pinyin</th>
<th>Phonetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>鱼 (yu)</td>
<td>fish</td>
<td>yu</td>
<td>/yu/</td>
</tr>
<tr>
<td>鸟 (niǎo)</td>
<td>bird</td>
<td>niǎo</td>
<td>/niǎo/</td>
</tr>
<tr>
<td>草 (cǎo)</td>
<td>grass</td>
<td>cǎo</td>
<td>/cao/</td>
</tr>
<tr>
<td>志 (zhì)</td>
<td>heart</td>
<td>zhì</td>
<td>/zhi/</td>
</tr>
</tbody>
</table>

- 鱼 (yu) = Fish - LI
- 鸟 (niǎo) = Bird - RA
- 草 (cǎo) = Vegetation - ZAO
- 志 (zhì) = Heart - SHI

**guó (country)** = ENCLOSURE - HUO
General operation: **SEMANTIC** - **PHONETIC**

<table>
<thead>
<tr>
<th>symbol</th>
<th>Pinyin</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>魚</td>
<td>yú</td>
<td>FLY</td>
</tr>
<tr>
<td>鳥</td>
<td>niǎo</td>
<td>BIRD</td>
</tr>
<tr>
<td>草</td>
<td>cǎo</td>
<td>GRASS</td>
</tr>
<tr>
<td>志</td>
<td>zhì</td>
<td>MIND</td>
</tr>
<tr>
<td>國</td>
<td>guó</td>
<td>COUNTRY</td>
</tr>
</tbody>
</table>

Knight/Sproat Writing Systems, Transliteration and Decipherment

General operation: **SEMANTIC** - **PHONETIC**

<table>
<thead>
<tr>
<th>symbol</th>
<th>Pinyin</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>魚</td>
<td>yú</td>
<td>FLY</td>
</tr>
<tr>
<td>鳥</td>
<td>niǎo</td>
<td>BIRD</td>
</tr>
<tr>
<td>草</td>
<td>cǎo</td>
<td>GRASS</td>
</tr>
<tr>
<td>志</td>
<td>zhì</td>
<td>MIND</td>
</tr>
<tr>
<td>國</td>
<td>guó</td>
<td>COUNTRY</td>
</tr>
</tbody>
</table>
A generalization of huìyì: Japanese kokuji (国字)

<table>
<thead>
<tr>
<th>Alex. #</th>
<th>Kokuj</th>
<th>Analysis</th>
<th>(Phonetic)</th>
<th>Kun</th>
<th>(on)</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>候</td>
<td>&lt;PERSON+MOVE&gt;</td>
<td>hataraki</td>
<td>do</td>
<td>‘effort’</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>風</td>
<td>&lt;WIND+STOP&gt;</td>
<td>nagi</td>
<td></td>
<td>‘lull, calm’</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>留</td>
<td>MOUNTAIN+UP+DOWN</td>
<td>tounge</td>
<td></td>
<td>‘mountain pass’</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>亜</td>
<td>&lt;HEART+FOREVER&gt;</td>
<td>kore</td>
<td></td>
<td>‘endure’</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>雨</td>
<td>&lt;FEW+HAIR&gt;</td>
<td>mushi</td>
<td></td>
<td>‘pluck’</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>倫</td>
<td>&lt;EAR+CERTAIN&gt;</td>
<td>shika</td>
<td></td>
<td>‘clearly’</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>婦</td>
<td>&lt;BODY+BEAUTIFUL&gt;</td>
<td>shitsuke</td>
<td></td>
<td>‘upbringing’</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>風</td>
<td>&lt;DOWN+WIND&gt;</td>
<td>oroshi</td>
<td></td>
<td>‘mountain wind’</td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>鳥</td>
<td>&lt;FIELD+BIRD&gt;</td>
<td>shig</td>
<td></td>
<td>‘snope’</td>
<td></td>
</tr>
<tr>
<td>249</td>
<td>娘</td>
<td>&lt;FEMALE+NOSÉ&gt;</td>
<td>kaki</td>
<td></td>
<td>‘wife’</td>
<td></td>
</tr>
<tr>
<td>138</td>
<td>薫</td>
<td>&lt;GRASS+ZA&gt;</td>
<td>薫 (on)</td>
<td>goza</td>
<td>‘matting’</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>鷺</td>
<td>&lt;TREE+MASA&gt;</td>
<td>鷺 (kun)</td>
<td>masa</td>
<td>‘straight grain’</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>釈</td>
<td>&lt;CLOTHING+YUKI&gt;</td>
<td>釈 (kun)</td>
<td>yuki</td>
<td>‘sleeve length’</td>
<td></td>
</tr>
</tbody>
</table>

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

‘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

- \(w\)h (verb 3-lit.) “set, place; add; stop; remain, last”; \(w\)h (adjective) “lasting”; \(w\)h jh “be patient” (literally, “lasting/set of heart”)
- \(w\)hyt (noun) “abundance (of grain)”
- \(w\)lj (noun) “columned hall” (literally, “marsh” of papyrus and lotus columns)
- \(w\)st (noun) “dominion”
- \(w\)st (noun) “Thebes” (nome and city)
- \(w\)s (noun) “ruin” (ininfinitive of 4ae-inf. verb \(w\)s “fall into ruin”)
- \(w\)tg (verb 4ae-inf.) “make festival”
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

<table>
<thead>
<tr>
<th>Sign sequence</th>
<th>Linear</th>
<th>Minoan</th>
<th>Greek</th>
<th>Greek</th>
<th>word meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tuninon</td>
<td>luninon</td>
<td>luninon</td>
<td>cumin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gunem</td>
<td>gunem</td>
<td>woman (gynecology)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>thnosos</td>
<td>thnosos</td>
<td>thnosos</td>
<td>odd (cornstarch)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>panem</td>
<td>panem</td>
<td>panem</td>
<td>father</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pharmakos</td>
<td>pharmakos</td>
<td>pharmakos</td>
<td>medicine (pharmacy)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teoys</td>
<td>teoys</td>
<td>teoys</td>
<td>so many</td>
<td></td>
</tr>
<tr>
<td></td>
<td>thoraks</td>
<td>thoraks</td>
<td>thoraks</td>
<td>thorax</td>
<td></td>
</tr>
<tr>
<td></td>
<td>quos</td>
<td>quos</td>
<td>quos</td>
<td>cow</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hippos</td>
<td>hippos</td>
<td>hippos</td>
<td>horse</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bixi</td>
<td>bixi</td>
<td>bixi</td>
<td>white (beechwood)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rhoa</td>
<td>rhoa</td>
<td>rhoa</td>
<td>rhino</td>
<td></td>
</tr>
</tbody>
</table>

### Cherokee (1821)

```
\begin{verbatim}
  Bole  awhup, kta, dsp ojebi p42i, tgz temep, dojaarei, p42i, p42i, tgz, ylp taoa, tb dojaarei, kta, dta ojebi p42i, dtaareez, tb yw dojaarei, qylat dsp ojebi p42i.
  Gwyz, ogr awhup, wpw, DSP, dSP ojebi p42i, tgz temep dojaarei, kta, dSP, OR, 0JSIB, dSP, TgZ, OJSIB, dSP ojebi p42i.
\end{verbatim}
```
Kana

万葉仮名 man'yōgana

Yi

Figure 2.13: Yi syllabary. http://www.omniajot.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

![Egyptian Consonantal Symbols](http://www.etymonline.com/images/writing/hieroglyphs.jpg)
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
Later Semitic scripts: vowel diacritics

The evolution of Greek writing

• Greek developed from Phoenician
• Vowel symbols developed by reinterpreting – or maybe misinterpreting – 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

<table>
<thead>
<tr>
<th>Catenation</th>
<th>Full form</th>
<th>Diacritic form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>अ &lt;a&gt;</td>
<td>क &lt;ka&gt;</td>
</tr>
<tr>
<td>आ &lt;aa&gt;</td>
<td>आ &lt;aa&gt;</td>
<td>क &lt;ka&gt;</td>
</tr>
<tr>
<td>ओ &lt;o&gt;</td>
<td>ओ &lt;o&gt;</td>
<td>क &lt;ko&gt;</td>
</tr>
<tr>
<td>ऑ &lt;au&gt;</td>
<td>ओ &lt;o&gt;</td>
<td>क &lt;kau&gt;</td>
</tr>
<tr>
<td>इ &lt;i&gt;</td>
<td>इ &lt;i&gt;</td>
<td>क &lt;kii&gt;</td>
</tr>
<tr>
<td>ए &lt;e&gt;</td>
<td>ए &lt;e&gt;</td>
<td>क &lt;ke&gt;</td>
</tr>
<tr>
<td>ऐ &lt;ai&gt;</td>
<td>ऐ &lt;ai&gt;</td>
<td>क &lt;kai&gt;</td>
</tr>
<tr>
<td>उ &lt;u&gt;</td>
<td>उ &lt;u&gt;</td>
<td>क &lt;ku&gt;</td>
</tr>
<tr>
<td>ऊ &lt;um&gt;</td>
<td>ऊ &lt;um&gt;</td>
<td>क &lt;kum&gt;</td>
</tr>
<tr>
<td>ऋ &lt;ri&gt;</td>
<td>ऋ &lt;ri&gt;</td>
<td>क &lt;kri&gt;</td>
</tr>
<tr>
<td>ऌ &lt;i&gt;</td>
<td>ऋ &lt;i&gt;</td>
<td>कि &lt;ki&gt;</td>
</tr>
</tbody>
</table>
### Kannada diacritic vowels

| Null | क | ₹
|------|---|---
| क़ | कू | का | कूँ |
| क़् | कृ | कृँ |
| क़्् | कृि | कृः |
| क़््् | कृिः |

- क़: Long "कू"
- क़्: Long "कृ"
- क़््: Long "कृि"
- क़्््: Long "कृिः"

\{i (.), LONG (铤)}: कू
\{e (.), uu (铤)}: कृ
\{e (.), ai (铤)}: कृि
\{e (.), LONG (铤)}: कृिः
\{e (.), uu (铤), LONG (铤)}: कृिः
Another alphasyllabary:
Ethiopic (Ge’ez) (4th century AD)
“Correct sounds for instructing the people”
(훈민정음)
The origin of Korean Writing

“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)
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”.
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

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

Knight/Sproat  Writing Systems, Transliteration and Decipherment  45

Text Elements       Characters

Composite:  Ç  Ç  Ç  Ç  Ç
Collation Unit:  ch  ch (Slovak)
Syllable:  क्षि  क्षि  क्षि  क्षि
Word:  cat  cat  cat  cat
### Example: Devanagari Code Points

![Devanagari Code Points]

---

**Figure 2-2. Characters Versus Glyphs**

<table>
<thead>
<tr>
<th>Glyphs</th>
<th>Unicode Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>त्र्र्र्र श्र</td>
<td>U+0923 LATIN CAPITAL LETTER A</td>
</tr>
</tbody>
</table>
Example of Logical Ordering: Tamil /hoo/

Q: Doesn’t the Tamil syllable பி பி பி 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.

B.

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

A.

B.
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

Gidi said, ""קִזְבּ תַּאָרְיָה בָּרְיָה"".
Unicode encoding schemes

Table 2-3. The Seven Unicode Encoding Schemes

<table>
<thead>
<tr>
<th>Encoding Scheme</th>
<th>Endian Order</th>
<th>BOM Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTF-8</td>
<td>N/A</td>
<td>yes</td>
</tr>
<tr>
<td>UTF-16</td>
<td>Big-endian or Little-endian</td>
<td>yes</td>
</tr>
<tr>
<td>UTF-16BE</td>
<td>Big-endian</td>
<td>no</td>
</tr>
<tr>
<td>UTF-16LE</td>
<td>Little-endian</td>
<td>no</td>
</tr>
<tr>
<td>UTF-32</td>
<td>Big-endian or Little-endian</td>
<td>yes</td>
</tr>
<tr>
<td>UTF-32BE</td>
<td>Big-endian</td>
<td>no</td>
</tr>
<tr>
<td>UTF-32LE</td>
<td>Little-endian</td>
<td>no</td>
</tr>
</tbody>
</table>

Figure 2-12. Unicode Encoding Schemes

UTF-32BE

UTF-32LE

UTF-16BE

UTF-16LE

UTF-8
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):

Knight/Sproat
Writing Systems, Transliteration and Decipherment

Phonemic transfer

Two spoken forms

When Languages Collide

At the border crossing (after writing):

Knight/Sproat
Writing Systems, Transliteration and Decipherment

Textual transfer

Two written forms
When Languages Collide

- Japanese/English example:
  KEVIN KNIGHT English writing
  K E V I N N A Y T English sounds
  K E B I N N A I T O Japanese sounds
  クビジト Japanese writing

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

- 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 …

Knight/Sproat Writing Systems, Transliteration and Decipherment
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)
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。

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アンダーで並んだ。

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ボギーで18位タイにつけている。
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

Knight/Sproat Writing Systems, Transliteration and Decipherment 67

A Model of Transcription

KEVIN KNIGHT
K E H V I H N N A Y T
K E B I N N A I T O
クビジン ナイト

English writing
English sounds
Japanese sounds

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

Angela Knight

[Knight & Graehl 98]
Spelling to Sound Transducer

- Richard talked about writing systems.
- Such a system captures an infinite relation of 
  \( \langle \text{sound-sequence}, \text{writing-sequence} \rangle \) pairs.

\[
\begin{align*}
\text{CAT} : & \quad \varepsilon \\
\varepsilon : & \quad K \\
\varepsilon : & \quad AE \\
\varepsilon : & \quad T \\
\text{SCAT} : & \quad \varepsilon \\
\varepsilon : & \quad S \\
\varepsilon : & \quad \varepsilon \\
\end{align*}
\]

\[
\text{WFST} \\
\text{words} \rightarrow \text{sounds} \\
\text{sounds} \rightarrow \text{words}
\]

Learning Sequence
Transformation Probabilities

Ideal training data:

\[
\begin{array}{cccc}
L & AE & M & P \\
\mid & \mid & \mid & \mid \\
r & a & n & p \\
\end{array}

\begin{array}{cccc}
S & T & IY & M \\
\mid & \mid & \mid & \mid \\
s & u & t & i & m & u \\
\end{array}
\]

\[
P(n \mid M) = 0.5 \\
P(m u \mid M) = 0.5
\]

need much more data, of course

Actual training data:

\[
\begin{array}{cccc}
L & AE & M & P \\
\mid & \mid & \mid & \mid \\
r & a & n & p & u \\
\end{array}

\begin{array}{cccc}
S & T & IY & M \\
\mid & \mid & \mid & \mid \\
s & u & t & i & m & u \\
\end{array}
\]

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]
A Model of Transcription

Angela Knight

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 K N I G H T English writing
KE VI N KN IGH T English letter chunks
クビン 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)

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] ...
Use of Transcription in Machine Translation Systems

Another approach [Hermjakob et al 08]

- Bilingual Training Corpus
- Transliteration Model
- New suggested phrasal translations (not mandatory use)
- Tagged test corpus
- Test Corpus
- MT system

Bilingual corpus, each side with transliterated items identified & marked

- Source side only, with transliterated items marked (throw away target side)

Trained monolingual “transliterate me” tagger (doesn’t just tag names!)

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

Cypriot Gekklish with Instant Messaging Shorthand:

<table>
<thead>
<tr>
<th>Cypriot Gekklish with Instant Messaging Shorthand:</th>
<th>ego n 3ero re pe8kia... skeftoume skeftoume omos tpt..</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized for automatic indexing or translation:</td>
<td>Εγώ εν ξέρω ρε παιθκία... σκέφτουμαι σκέφτουμαι όμως τίποτα...</td>
</tr>
</tbody>
</table>

see "Greeklish", Wikipedia
Overview of the Transliteration/Transcription Literature

We have only touched on what is a large literature.


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
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.

<table>
<thead>
<tr>
<th>English NE</th>
<th>Russian NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lilic</td>
<td>лилия</td>
</tr>
<tr>
<td>fletcher</td>
<td>флетчер</td>
</tr>
<tr>
<td>bradford</td>
<td>брандфорд</td>
</tr>
<tr>
<td>isabel</td>
<td>изабель</td>
</tr>
<tr>
<td>hoffmann</td>
<td>гофман</td>
</tr>
<tr>
<td>kathmandu</td>
<td>катманду</td>
</tr>
</tbody>
</table>

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, $(E_S, E_T) = (\text{powell}, \text{pouel})$, $n = 2$
    - $E_S \rightarrow \{ _, p, o, w, e, l, po, ow, we, el, ll \}$
    - $E_T \rightarrow \{ _, p, o, u, e, l, po, ou, ue, el \}$
    - Feature vector is thus $((p, _), (p, a), \ldots (w, au), \ldots (el, el), \ldots (ll, el))$

- 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**

$D \leftarrow \emptyset$

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
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

<table>
<thead>
<tr>
<th>Iteration 0</th>
<th>Iteration 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 скоро {с, ск, ску, скр}</td>
<td>1 форсайт {ф, фо, фоу, фоу}</td>
</tr>
<tr>
<td>2 офоры {о, ор, ори, орэ}</td>
<td>2 офоры {о, ор, ори, орэ}</td>
</tr>
<tr>
<td>3 кокрын {к, кр}</td>
<td>3 проры {пр, прэ}</td>
</tr>
<tr>
<td>4 флоры {ф, фо, фоу}</td>
<td>4 фрос</td>
</tr>
<tr>
<td>24 форсайт {ф, фо, фоу}</td>
<td></td>
</tr>
</tbody>
</table>

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 Aššurnāṣirpal II
Not every communicative symbol system is writing

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

In Figure 4.12, the name of Ptolemy is written in Egyptian:

\[ ptwlm\text{ys} \]

Figure 4.12: *Ptolemy* from the Rosetta stone.

In Figure 4.13, the name of Cleopatra is written in Egyptian:

\[ l\text{y}w\text{r} \]

Figure 4.13: *Cleopatra* from the Bankes Obelisk.

The name *Ptolemy* is written as \( ptwlm\text{ys} \) (\( \Pi\tau\Omega\Lambda\varepsilon\text{MAIO}Σ \)), which means *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 (\( gǔsītāfū \))
- 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 guessed that vertical bars in the inscription were word separators.
  - By the large number of symbols between the separators he reasoned that the system must be alphabet. (Actually that turned out to 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 *xšeršē* and *darheuš*
  - The later word for ‘king’ was *kšēio*
- 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

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
  - ru ki ti jo
  - ru ki ti ja
  - a mi ni so
  - a mi ni si jo
  - 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 (*iqqwos, later hippos ‘horse’)

- No definite articles

- Confirmation from new finds by Carl Blegen:
  - ti ri po de
  - qe to ro we
  - qʷetrōwes
  - 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.

Rawlinson: Then I went on to the country of Comukha, which was disobedient and withheld 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 Dummuhi, the enemy who owed tribute and gifts to the god Asur, my lord.
Verification

A text from Abu Simbel

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

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

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 RDCLIO JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”

KDCY LQZKTLJQX CY MDBCYJQOL: “TR
HYD FKXC, FQ MKX RLQQIQ HYDL
MKL DXCTW RDCLIO JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”
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.”

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

3-gram letter-based language model of English used for decipherment

Unknown Script as a Cipher

Greek sounds

Modern Mayan sounds

“make the text speak”

(Roman tablet)

(Mayan writing)
Unknown Script as a Cipher

ciphertext (6980 letters)

primera parte

del ingenioso

hidalgo don ...

(Don Quixote)

Modern
Spanish
sounds

26 sounds:
B, D, G, J (canyon),
L (yarn), T (thin), a,
b, d, e, f, g, i, k, l,
m, n, o, p, r,
rr (trilled), s,
t, tS, u, x (hat)

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

ciphertext (6980 letters)

primera parte

del ingenioso

hidalgo don ...

(Don Quixote)

Modern
Spanish
sounds

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

Phoneme bigram model
P(L | tS) = 0.003

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

P(c | p) =
P(c1 | p1) * 
P(c2 | p2) * 
P(c3 | p3) * …
EM-based decipherment finds a very good “key” and achieves 93% phoneme accuracy

Correct sounds:    \textit{primera parte del inxenioso iDalGo don kixote...}
Deciphered sounds: \textit{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 $P(c \mid p)$ 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
  - $\{stdbptk\}$ $\{mnrl\}$ $\{V\}$ $\{mnrl\}$ $\{stdbptk\}$
- Physiological preference constraints
Universal Phonology

Task 1: Label each letter with a phoneme

human sounding sequence

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

# of syllables in word syllable type sequence consonant/vowel sequences

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 vccvcvccv 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
Azolzértik gabe ez zaio inori bere jakbegaak kenduko
Den ebet ne vo tennet e bercherenèsh digantañ diouzh s’hoant
Нищий не требуется да бъде произволно лишён от своего собственности
Нинъо no serà privat arbitráriament de la seva propietat
Ди а со прупиïтă ун пӧ эssa приuu ниму ді моду тиранніцу
Нитко не сумне самовольно біти лишень своє имовине
Niko nesmi být svěvověně zbaven svého majetku
Инге маг вилкіріт баревес син еджемон
Ниemand mag willekeurig van zijn eigendom worden beroofd

Nul ne peut être arbitrairement privé de sa propriété
Nimmer mei samar fan syn eigendom berøve wurde
Ningiu serà privado arbitrariamente da sua proprietà
Niemand darf willkürlich seines Eigentums beraubt werden
Kovaçi është u brendi që kërkohet nga dëshmitë e tij
Аваалгүүнгүнүн дөө жана классикалык ишбир эскеруу
Дэн ага ноттэ ки дукийыр жана табак дарык ар көмөк
Ден ебет акын бо ахай эмс көлбөлөк аймаак
den ebet ne vo tennet e bercherenèsh digantañ diouzh s’hoant
Ba wanda za a kwace wa dukiyyara ba tare da cikakken dalli ba
Senkit sem leheb tulajdonatol õnkin te lõndlõ stani
Engan má efir geðþótta svipta eign sinn
Nesuno essera privata arbitrariamente de su proprietate
Ní féidir a bhainin a bhaint go forlámhach de dhúine ar bith
Al neni u este arbitr forprenita lia proprieto
Keltälti ei tohi lema vara meelevaldsett ára vótta
Eingin skal hissini vera fyrir ongarto
Me kua ni dua e kovei vua na nona iyau
Keltään älköön mielivaltaisesti riistättäkö hänen omaisuuttaan
**Unknown Source Language**

- **Input:**
  
  `cevzren cnegr qry vatrabfb uvqnytb qba dhvwbgr qr ynznapun ...`

- **Top 5 languages with best P(c) after deciphering:**
  
<table>
<thead>
<tr>
<th>P(c)</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.29120</td>
<td>spanish</td>
</tr>
<tr>
<td>-5.43346</td>
<td>galician</td>
</tr>
<tr>
<td>-5.44087</td>
<td>portuguese</td>
</tr>
<tr>
<td>-5.48023</td>
<td>kurdish</td>
</tr>
<tr>
<td>-5.49751</td>
<td>romanian</td>
</tr>
</tbody>
</table>

- **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 asequ ...`

- **Simultaneous language ID and decipherment**

**Transliteration as a Cipher**

- **Ciphertext:** Japanese Katakana
- **Plaintext:** English
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

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


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.

Further Reading

Decipherment
Further Reading

Auto Decipherment


