Information was passed between our ancestors first through genes, then gestures, then speech, then drawings.
Imagine your ancestor wanted to leave the message “there are ox halfway up the river”
Ancient Egyptian (c. 3000 BCE)
- Few writers
- Stone tablets
- Many (>1500) symbols representing ideas (e.g., apple)
- A few (~140) symbols representing sounds (e.g., gah)

Demotic (c. 650 BCE)
- Many writers
- Papyrus sheets
- More purposes (e.g., recipes, contracts)
- Fewer symbols
- Higher proportion of symbols representing sounds
The Rosetta stone

- The **Rosetta stone** dates from 196 BCE.
- It was re-discovered by French soldiers during Napoleon’s invasion of Egypt in 1799 CE.

- It contains three **parallel** texts in different languages, only the **last** of which was understood.
- **By 1799, ancient Egyptian had been forgotten.**
Writing systems

• **Logographic**: *adj.* Describes writing systems whose symbols denote **semantic** ideas.

• **Phonographic**: *adj.* Describes writing systems whose symbols denote **sounds**. E.g., in English the symbols ‘sh’ mean...

• Some writing systems are a mix of these qualities:
  • 媽 mā ‘mother’, formed from:
  • 女 nǚ (means like) ‘woman’
  • 馬 mǎ (sounds like) ‘horse’
Writing systems

- **Logographic**: Symbols refer to ideas.
- **Phonographic**: Symbols refer to sounds.

English carries logographic heritage.

```
Proto-Sinaitic

Phoenician

Cyrillic
```

Is ancient Egyptian logographic or phonographic?
Deciphering Rosetta

- During 1822–1824, Jean-François Champollion worked on the Rosetta stone. He noticed:
  1. The circled Egyptian symbols appeared in roughly the same positions as the word ‘Ptolemy’ in the Greek.
  2. The number of Egyptian hieroglyph tokens were much larger than the number of Greek words → Egyptian seemed to have been partially phonographic.
  3. Cleopatra’s cartouche was written
Aside – deciphering Rosetta

- So if PTOLEMY was ‘Ptolemy’ and CLEOPATRA was ‘Cleopatra’ and the symbols corresponded to sounds – can we match up the symbols?

<table>
<thead>
<tr>
<th>P</th>
<th>T</th>
<th>O</th>
<th>L</th>
<th>M</th>
<th>E</th>
<th>Y</th>
</tr>
</thead>
</table>
| C | L | E | O | P | A | T | R | A

- This approach demonstrated the value of working from parallel texts to decipher an unknown language:
  - *It would not have been possible without aligning unknown words (hieroglyphs) to known words (Greek)*...
Today

• Introduction to statistical machine translation (SMT).
  • What we want is a system to take utterances/sentences in one language and transform them to another:

  Ne mange pas ce chat!

  Don’t eat that cat!
Transliteration

- A bilingual dictionary that aligns words across languages can be helpful, but only for simple cases.

<table>
<thead>
<tr>
<th>¿</th>
<th>Dónde</th>
<th>está</th>
<th>la</th>
<th>biblioteca</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where</td>
<td>is</td>
<td>the</td>
<td>library</td>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>

| Où | est | la | bibliothèque | ? |

<table>
<thead>
<tr>
<th>Mi</th>
<th>nombre</th>
<th>es</th>
<th>T-bone</th>
</tr>
</thead>
<tbody>
<tr>
<td>My</td>
<td>name</td>
<td>is</td>
<td>T-bone</td>
</tr>
</tbody>
</table>

| Mon | nom | est | T-bone |

CSC401/2511 – Spring 2015
Challenge 1: lexical ambiguity

A word token in one language may have many possible translations in another:

- E.g., book the flight → reservar
  read the book → libro

  the chair in the chair → président, chaise

  kill the queen → tuer la reine
  kill the Queen → éteindre le music de Queen
Challenge 2: differing word orders

- **English**: subject – (trans.) verb – object
  - **Japanese**: subject – object – (trans.) verb

  e.g., **English**: IBM bought Lotus
  - **Japanese**: ~IBM Lotus bought

- **English**: determiner – adjective – noun
  - **French**: determiner – noun – adjective

  e.g., **English**: the fast zombie
  - **French**: le zombie rapide
Challenge 3: unpreserved syntax

- Differences in syntax between languages are felt over longer distances than simple word alternations.
  - E.g.,

    La botella entró a la cuerva flotando
    (the bottle entered to the cave floating)

    The bottle floated into the cave

- This implies that we’d need high-level grammars of the source and target languages.
Challenge 4: syntactic ambiguity

• **Syntactic ambiguity** in the source makes it difficult to produce a single sentence in the target language.
  
  • *E.g.*, 

  **Rick hit the zombie with the stick**

  - *Rick golpeó el zombie con el palo*  
    (the stick was used)
  - *Rick golpeó el zombie que tenía el palo*  
    (the zombie had the stick)
Challenge 5: idiosyncracies

• Languages have their own idioms, and “feel”.
  • E.g.,

  We have to burn the midnight oil
  Il faut travailler tard
  Il faut brûler l’huile de minuit

  Estie de sacramouille
  Host of the sacrament
  By golly!
Classical MT: Dictionaries

• Early MT involved merely looking up each word in a bilingual dictionary of rules.
• E.g., translate ‘much’ or ‘many’ into Russian:

If preceding word is how return skol’ko
else if preceding word is as return stol’ko zhe
else if word is much
    if preceding word is very return nil
    else if following word is a noun return mnogo
else (word is many)
    if preceding word is a preposition and next word is a noun return mnogii
else return mnogo

From Jurafsky & Martin
Classical MT: Dictionaries

• This approach causes some problems, e.g.,

• It’s difficult/impossible to capture long-range re-orderings:
  • English: Sources said that IBM bought Lotus yesterday
  Japanese: ~Sources yesterday IBM Lotus bought that said

• It’s difficult to disambiguate parts-of-speech:
  • English: They said that I punched that zombie
  • French: Ils ont dit que j'ai frappé ce zombie

• Having experts write lots of rules can become unruly.
  • ...and expensive...and full of mistakes...
Classical MT: Transfer-based approach

• **Transfer-based** MT involves three phases:

  • **Analysis:** e.g., build *syntactic parse trees* of the source sentence.
  • **Transfer:** e.g., convert the *source*-language parse tree to a *target*-language parse tree.
  • **Generation:** e.g., produce an *output sentence* from the target-language parse tree.

• These systems can involve fairly deep analysis, often including *semantic* analysis.
Example of syntactic transfer

See csc485/2501 for more on computational approaches to parse trees

From Regina Barzilay at MIT
Example of syntactic transfer

Transformations are defined at the syntactic level

From Regina Barzilay at MIT
Classical MT: Transfer-based approach

• Transferring between parse trees allows us to encode more general rules with long-term dependencies.

• However, if we want to translate between $L$ languages, we’d need $O(L^2)$ sets of transformation rules.
  • This would involve lots of experts in each language ($\$$).
  • This can be somewhat mitigated by abstracting beyond syntax into an interlingua: a conceptual space common to all languages.
    • We might need a workable theory of neurolinguistics to do this properly, but ‘hacks’ are getting some good results.
Statistical machine translation

- Machine translation seemed to be an intractable problem by the late 1940s until a change in perspective...

When I look at an article in Russian, I say: ‘This is really written in English, but it has been **coded** in some strange symbols. I will now proceed to **decode**.’

Warren Weaver  March, 1947

Claude Shannon  July, 1948
How not to use the noisy channel

• The model $P(E, F)$ tells us how likely an English sentence $E$ and a French sentence $F$ are to correspond to each other.

• Imagine that you’re given a French sentence, $F$, and you want to convert it to the best corresponding English sentence, $E^*$
  • i.e.,
    $$E^* = \arg\max_{E} P(E, F)$$

• Others may be tempted to model this as
  $$E^* = \arg\max_{E} P(E|F)P(F)$$
  This is useless if you’re always given $F$
How not to use the noisy channel

• Others may be tempted to model this as

\[ E^* = \underset{E}{\text{argmax}} \; P(E|F)P(F) \]

This is useless if you’re always given \( F \)

• If \( P(E|F) \) is a model that translates word-to-word, then we cannot account for differing word orders across languages.
  • E.g., Source French: \( le \) \( zombie \) \( rapide \)
    Target English: \( the \) \( zombie \) \( fast \)

• If \( P(E|F) \) includes syntax, it becomes very difficult to learn without experts or specially-annotated data.
The noisy channel

\[
E^* = \arg \max_E P(F|E)P(E)
\]
How to use the noisy channel

• How does this work?

\[ E^* = \operatorname{argmax}_E P(F|E)P(E) \]

• \( P(E) \) is a language model (e.g., N-gram) and encodes knowledge of word order.

• \( P(F|E) \) is a word-level translation model that encodes only knowledge on an unordered word-by-word basis.

• Combining these models can give us naturalness and fidelity, respectively.
How to use the noisy channel

• Example from Koehn and Knight using only conditional likelihoods of Spanish words given English words.

• Que hambre tengo yo
  \[ \rightarrow \]
  What hunger have I \( P(S|E) = 1.4E^{-5} \)
  Hungry I am so \( P(S|E) = 1.0E^{-6} \)
  I am so hungry \( P(S|E) = 1.0E^{-6} \)
  Have I that hunger \( P(S|E) = 2.0E^{-5} \)
  ...

How to use the noisy channel

• ... and with the English language model

• Que hambre tengo yo
  →
  What hunger have I \( P(S|E)P(E) = 1.4E^{-5} \times 1.0E^{-6} \)
  Hungry I am so \( P(S|E)P(E) = 1.0E^{-6} \times 1.4E^{-6} \)
  I am so hungry \( P(S|E)P(E) = 1.0E^{-6} \times 1.0E^{-4} \)
  Have I that hunger \( P(S|E)P(E) = 2.0E^{-5} \times 9.8E^{-7} \)
  ...

How to learn $P(F|E)$?

• Solution: collect statistics on vast parallel texts

  e.g., the *Canadian Hansards*: bilingual Parliamentary proceedings

  ... *citizen* of Canada has the *right* to vote in an election of members of the House of Commons or of a legislative assembly and to be qualified for membership ...

  ... *citoyen* canadien a le *droit* de vote et est éligible aux élections législatives fédérales ou provinciales ...
Bilingual data

Data from Linguistic Data Consortium at University of Pennsylvania.

From Chris Manning’s course at Stanford
Alignment

• In practice, words and phrases can be out of order.

According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above average growth rates.

Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d’adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

From Manning & Schütze
Alignment

- Also in practice, we’re usually not given the alignment.

According to our survey 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above average growth rates.

Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d’adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

From Manning & Schütze
Sentence alignment

• Sentences can also be unaligned across translations.

• E.g.,  

  *He was happy.*  
  
  \[ \rightarrow \]  
  
  *He had bacon.*  

  *Il était heureux parce qu'il avait du bacon.*

\[
\begin{align*}
E_1 & \quad F_1 \\
E_2 & \quad F_2 \\
E_3 & \quad F_3 \\
E_4 & \quad F_4 \\
E_5 & \quad F_5 \\
E_6 & \quad F_6 \\
E_7 & \quad F_7 \\
\ldots & \quad \ldots
\end{align*}
\]
Sentence alignment

• We often need to align sentences before we can align words.

• We’ll look at two broad classes of methods:
  1. Methods that only look at sentence length,
  2. Methods based on lexical matches, or “cognates”.
1. Sentence alignment by length

(Gale and Church, 1993)

- Assuming the paragraph alignment is known,
  - $L_E$ is the # of words in an English sentence,
  - $L_F$ is the # of words in a French sentence.

- Assume $L_E$ and $L_F$ have Gaussian/normal distributions with $\mu = cL_X$ and $\sigma^2 = s^2L_X$.
  - Empirical constants $c$ and $s$ set ‘by hand’.
  - The penalty, $Cost(L_E, L_F)$, of aligning sentences with different lengths is based on the divergence of these Gaussians.
1. Sentence alignment by length

We can associate costs with different types of alignments. \( C_{i,j} \) is the prior cost of aligning \( i \) sentences to \( j \) sentences.

\[
\text{Cost} = \text{Cost}(\mathcal{L}_{E_1} + \mathcal{L}_{E_2}, \mathcal{L}_{F_1}) + C_{2,1} + \\
\text{Cost}(\mathcal{L}_{E_3}, \mathcal{L}_{F_2}) + C_{1,1} + \\
\text{Cost}(\mathcal{L}_{E_4}, \mathcal{L}_{F_3}) + C_{1,1} + \\
\text{Cost}(\mathcal{L}_{E_5}, \mathcal{L}_{F_4} + \mathcal{L}_{F_5}) + C_{1,2} + \\
\text{Cost}(\mathcal{L}_{E_6}, \mathcal{L}_{F_6}) + C_{1,1}
\]

Find distribution of sentence breaks with minimum cost using dynamic programming.
2. Sentence alignment by cognates

- **Cognates**: *n.pl.* Words that have a common **etymological** origin.
- **Etymological**: *adj.* Pertaining to the historical derivation of a word. E.g., *porc*→*pork*

- The intuition is that words that are **related** across languages have similar **spellings**.
  - e.g., *zombie/zombie, government/gouvernement*
  - Not always: *son* (male offspring) vs. *son* (sound)

- Cognates can “anchor” sentence alignments between related languages.
2. Sentence alignment by cognates

• Cognates should be spelled similarly...

• **N-graph**: *n.* Similar to *N*-grams, but computed at the **character-level**, rather than at the word-level.
  
  E.g., $\text{Count}(s, h, i)$ is a **trigraph** model

• Church (1993) tracks all **4-graphs** which are identical across two texts.
  
  • He calls this a ‘signal-based’ approximation to cognate identification.
2a. Church’s method

1. Concatenate paired texts.

2. Place a ‘dot’ where the $i^{th}$ French and the $j^{th}$ English 4-graph are equal.

3. Search for a short path ‘near’ the bilingual diagonals.

From Manning & Schütze

E.g., the $i^{th}$ French 4-graph is equal to the $j^{th}$ English 4-graph.
2a. Church’s method

- Each point along this path is considered to represent a match between languages.
- The relevant English and French sentences are aligned.

From Manning & Schütze

E.g., the \( p^{th} \) French sentence is aligned to the \( q^{th} \) English sentence.
2b. Melamed’s method

- \(LCS(A, B)\) is the **longest common subsequence** of characters (with gaps allowed) in words \(A\) and \(B\).

- Melamed (1993) measures similarity of words \(A\) and \(B\)

\[
LCSR(A, B) = \frac{\text{length}(LCS(A, B))}{\max(\text{length}(A), \text{length}(B))}
\]

- e.g.,

\[
LCSR(\text{government}, \text{gouvernement}) = \frac{10}{12}
\]
2b. Melamed’s method

- Excludes **stop words** from both languages.
  (e.g., *the*, *a*, *le*, *un*)

- Melamed empirically declared that cognates occur when $LCSR \geq 0.58$ (i.e., there’s a lot of overlap in those words).
  - $\therefore$ 25% of words in Canadian Hansard are cognates.

- As with Church, construct a “bitext” **graph**.
  - Put a point at position $(i, j) \iff LCSR(i, j) \geq 0.58$.
  - Find a near-diagonal alignment, as before.
From sentences to words

• We’ve computed the **sentence** alignments.

• What about **word** alignments?
Word alignment

- **Word alignments** can be 1:1, N:1, 1:N, 0:1, 1:0, ... E.g.,
  - “zero fertility” word: not translated (1:0)
  - “spurious” words: generated from ‘nothing’ (0:1)

**Note**: that this is only one **possible** alignment.

One word translated as several words (1:N)
Intuition of statistical MT

• The words ‘the’ and ‘maison’ co-occur frequently, but not as frequently as ‘the’ and ‘la’.

\[ P(\text{la}|\text{the}) \] should be higher than \[ P(\text{fleur}|\text{the}) \], \[ P(\text{bleue}|\text{the}) \], and even \[ P(\text{maison}|\text{the}) \]

Note: we consider all possible word alignments....
Assignment 2 – content

• **Build** $N$-gram language models, with smoothing.

• **Learn** word-level alignments with the IBM-1 model using data from the Canadian Hansard.

• **Combine** the language and alignment models into a simple French-to-English translator.

• There are some **bonus** marks available for substantially going beyond the minimal requirements.
Assignment 2 – languages

• Sentences have already been **split** and **aligned** for you.
  • Words have **not** been aligned.

• You **don’t** need to know French for this assignment.
  • French is more ‘rigid’ than English, so its use of contractions, e.g., are more **regular**.
  • You have to do some pre-processing of French sentences, but those rules are given to you **explicitly**.
Assignment 2 – practical

• Will be posted by Friday 6 February.

• Will be programmed in Matlab.
  • Various support functions for this assignment will be available on CDF.
  • Patricia A. Thaine will give a tutorial on Matlab on 13 February.

• Marks will be given more for understanding the algorithms and concepts than for specific results.
### Puzzle: Machine translation

- **Puzzle *(for fun)*:** Translate this Centauri phrase to Arcturan: “farok crrrok hihok yorok clok kantok ok-yurp”

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1a.</strong> ok-voon ororok sprok .</td>
<td><strong>7a.</strong> lalok farok ororok lalok sprok izok enemok .</td>
<td><strong>7b.</strong> wat jjat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td><strong>1b.</strong> at-voon bichat dat .</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2a.</strong> ok-drubel ok-voon anok plok sprok .</td>
<td><strong>8a.</strong> lalok brok anok plok nok .</td>
<td><strong>8b.</strong> iat lat pippat rrat nnat .</td>
</tr>
<tr>
<td><strong>2b.</strong> at-drubel at-voon pippat rrat dat .</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3a.</strong> erok sprok izok hihok ghirok .</td>
<td><strong>9a.</strong> wiwok nok izok kantok ok-yurp .</td>
<td><strong>9b.</strong> totat nnatquat oloat at-yurp .</td>
</tr>
<tr>
<td><strong>3b.</strong> totat dat arrat vat hilat .</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4a.</strong> ok-voon anok drok brok jok .</td>
<td><strong>10a.</strong> lalok mok nok yorok ghirok clok .</td>
<td><strong>10b.</strong> wat nnat gat mat bat hilat .</td>
</tr>
<tr>
<td><strong>4b.</strong> at-voon krat pippat sat lat .</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5a.</strong> wiwok farok izok stok .</td>
<td><strong>11a.</strong> lalok nok crrrok hihok yorok zanzanok .</td>
<td><strong>11b.</strong> wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td><strong>5b.</strong> totat jjatquat cat .</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>6a.</strong> lalok sprok izok jok stok .</td>
<td><strong>12a.</strong> lalok rarok nok izok hihok mok .</td>
<td><strong>12b.</strong> wat nnat forat arrat vat gat .</td>
</tr>
<tr>
<td><strong>6b.</strong> wat dat kratquat cat .</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Puzzle: Machine translation

**Hint to get started:**

“farok crrrok hihok yorok clok kantok ok-yurp”

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok farok ororok lalok sprok izok enemok .</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lalok brok anok plok nok .</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat rrat dat .</td>
<td>8b. iat lat pippat rrat nnat .</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat nnat quat oloat at-yurp .</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
</tr>
<tr>
<td>4b. at-voon krat pippat rrat dat .</td>
<td>10b. wat nnat gat mat bat hilat .</td>
</tr>
<tr>
<td>5a. wiwok farok izok stok .</td>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
</tr>
<tr>
<td>5b. totat jjat quat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok rarok nok izok hihok mok .</td>
</tr>
<tr>
<td>6b. wat dat krat quat cat .</td>
<td>12b. wat nnat forat arrat vat gat .</td>
</tr>
</tbody>
</table>
Readings

• Manning & Schütze: Sections 13.0 and 13.2

• (optional) Gale & Church “A Program for Aligning Sentences in Bilingual Corpora” (on course website)