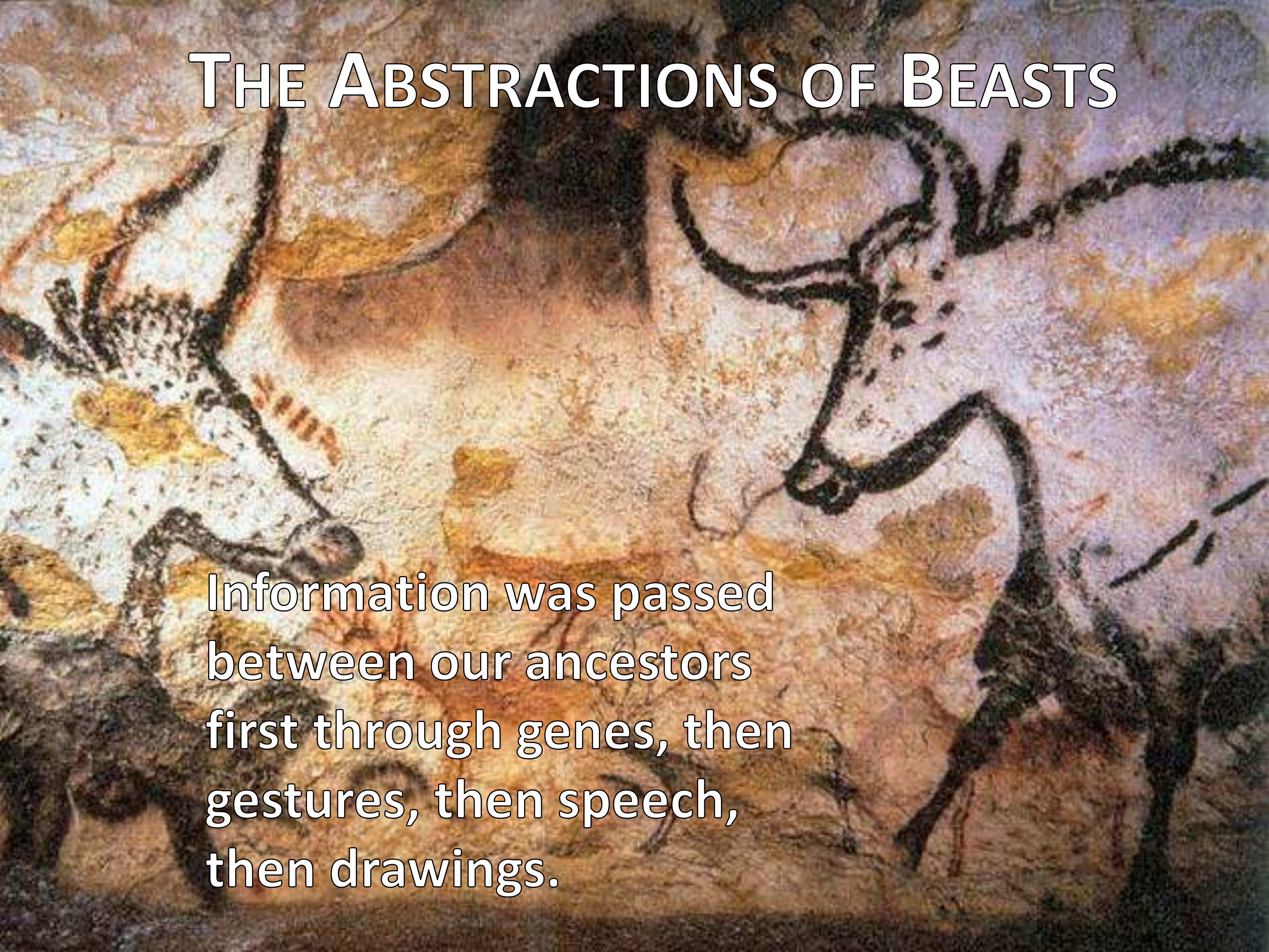
statistical machine translation

Part 1: Introduction & Sentence Alignment

CSC401/2511 – Natural Language Computing – Spring 2019
Lecture 6 Frank Rudzicz and Chloé Pou-Prom
University of Toronto

Statistical Machine Translation

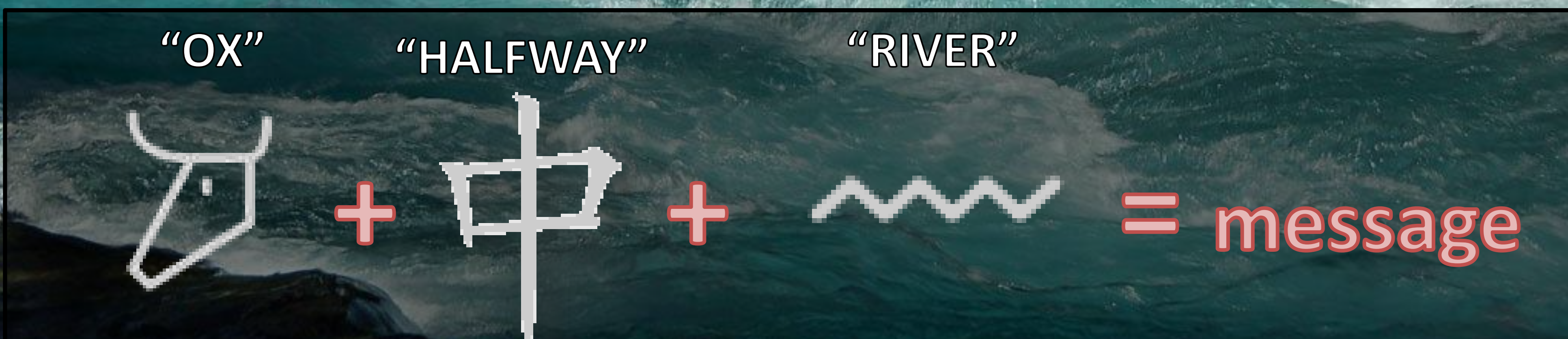
- Challenges to statistical machine translation
- Sentence alignment
- IBM model
- Phrase-based translation
- Decoding
- Evaluation



The Abstractions of Beasts

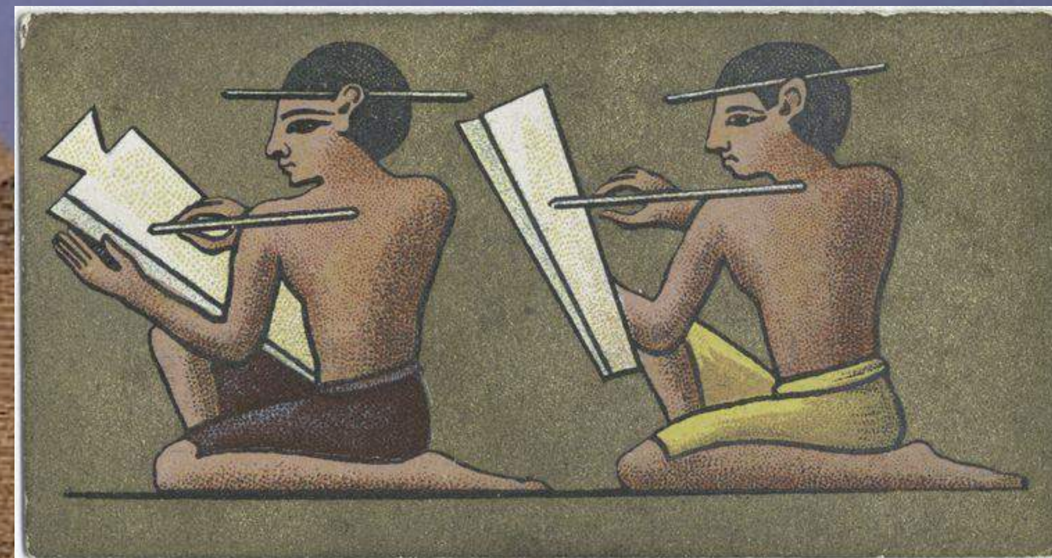
Information was passed between our ancestors first through genes, then gestures, then speech, then drawings.

Imagine your ancestor wanted to leave the message *"there are ox halfway up the river"*



IDEOGRAM

PICTOGRAMS



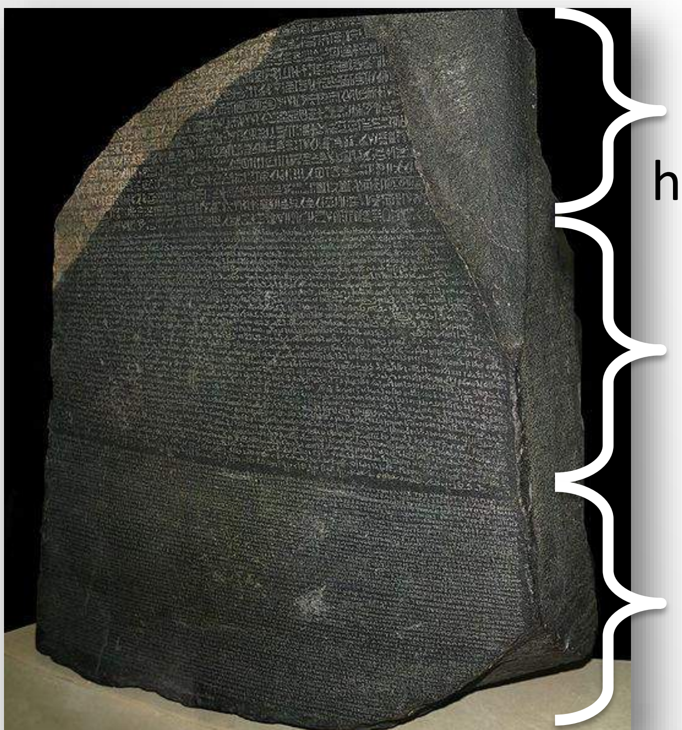
Ancient Egyptian (c. 3000 BCE)

- **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.



Ancient Egyptian hieroglyphs

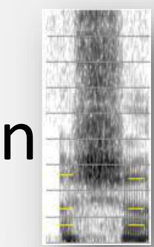
Egyptian Demotic

Ancient Greek

- It contains three **parallel** texts in different languages, only the **last** of which was understood.

- *By 1799, ancient Egyptian had been forgotten.*

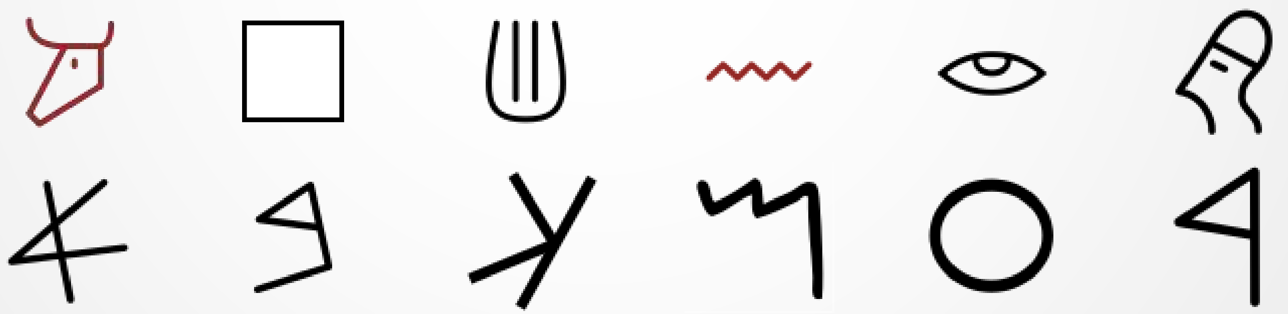
Writing systems

- **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.

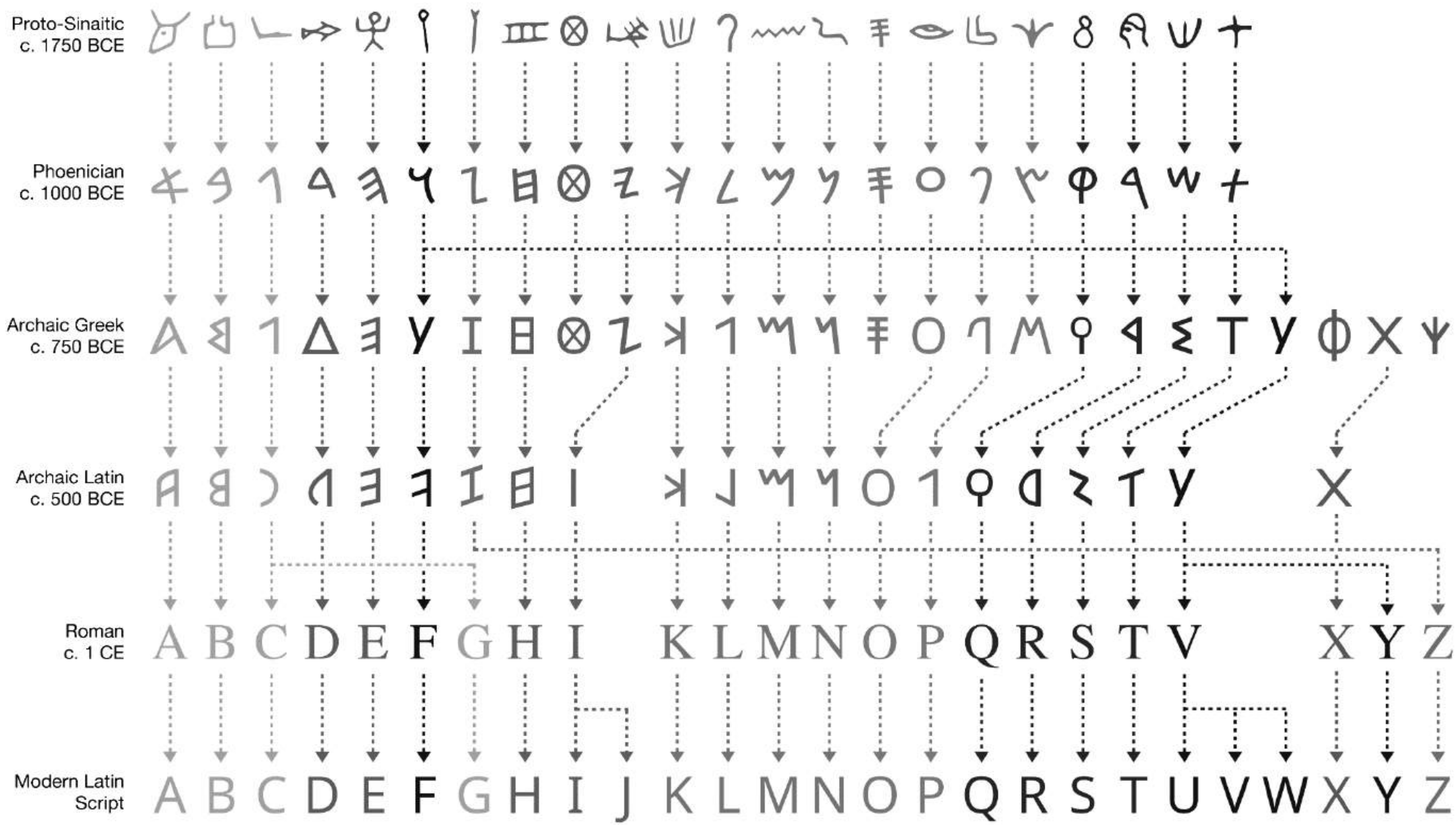


	"alph" (ox)	"bet" (house)	"kaf" (palm)	"mem" (water)	"en" (eye)	"ro" (head)
Proto-Sinaitic						
Phoenician						
Cyrillic	A	b	K	M	O	P

Is ancient Egyptian logographic or phonographic?

UNIVERSITY OF TORONTO

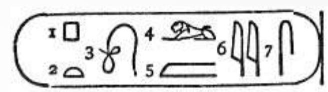
Evolution of the Alphabet



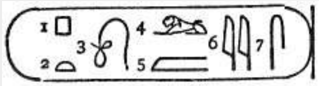
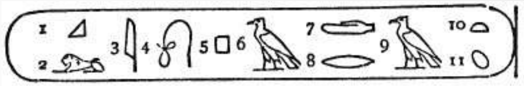
By Matt Baker | UsefulCharts.com

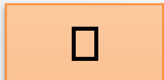
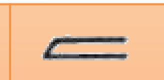
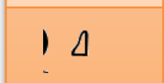
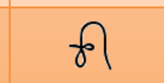
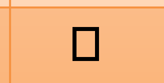
Alphabetum font by Juan José Marcos used for Ancient Greek/Latin

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

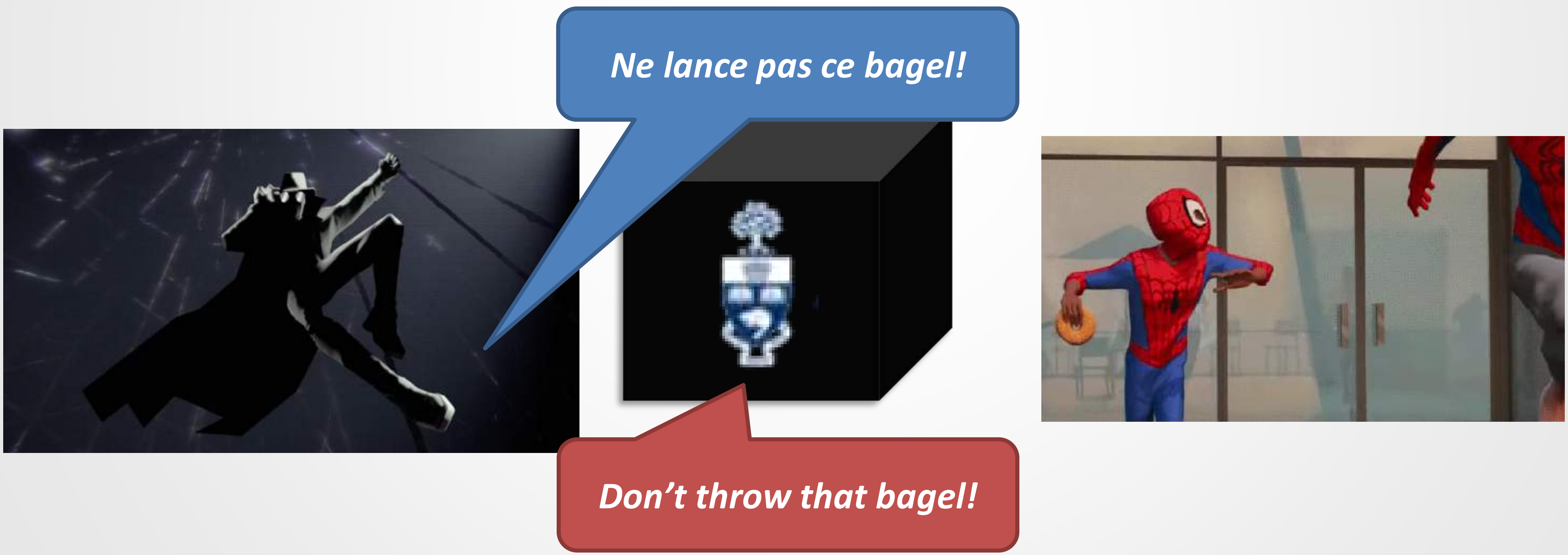
- So if  was '*Ptolemy*' and  was '*Cleopatra*' and the symbols corresponded to sounds – can we match up the symbols?

								
P	T	O	L	M	E	Y		
								
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 (hieroglyhs) 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:



Direct translation

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

¿	*Dónde*	*está*	*la*	*biblioteca*	*?*
	Where	*is*	*the*	*library*	*?*
	Où	*est*	*la*	*bibliothèque*	*?*

Mi	*nombre*	*es*	*T-bone*
My	*name*	*is*	*T-bone*
Mon	*nom*	*est*	*T-bone*

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 la musique 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***)

↓

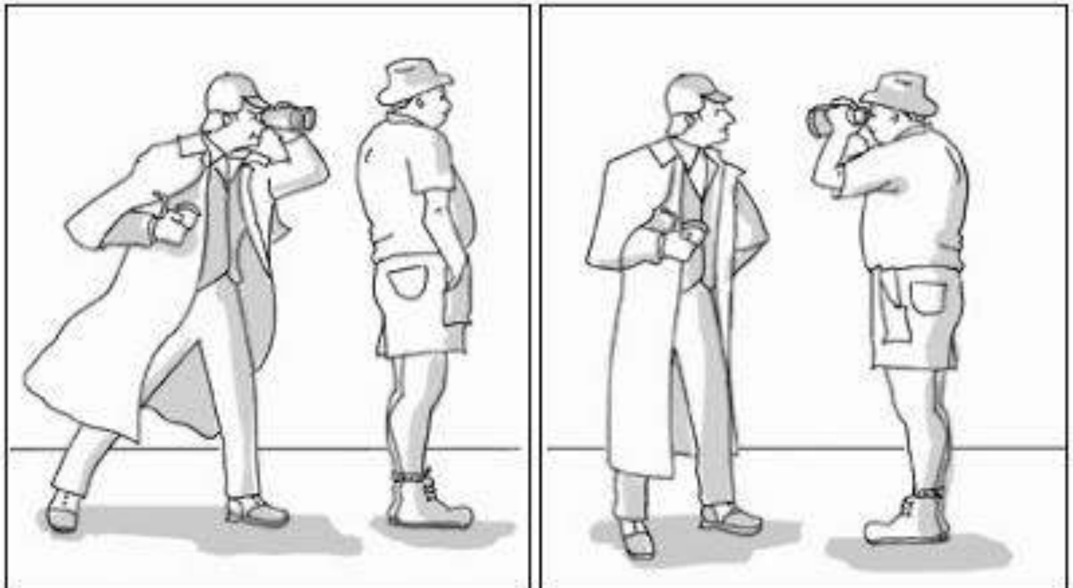
*The bottle **floated** into the cave*

- 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.,*

Sherlock saw the man using binoculars



Sherlock saw the man using binoculars. Sherlock saw the man using binoculars.

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 Morty with the stick

Rick golpeó el Morty con el palo
(the stick was used)

Rick golpeó el Morty que tenia el palo
(the Morty 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! ✓

L'eau dans la cave → *Water in the basement* ✗

Your pants are short ✓

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 **<u>that</u>** I punched **<u>that</u>** Morty*
    - French: *Ils ont dit **<u>que</u>** j'ai frappé **<u>ce</u>** Morty*
  - Having experts write lots of rules can become unruly.
    - …and expensive…and full of mistakes…

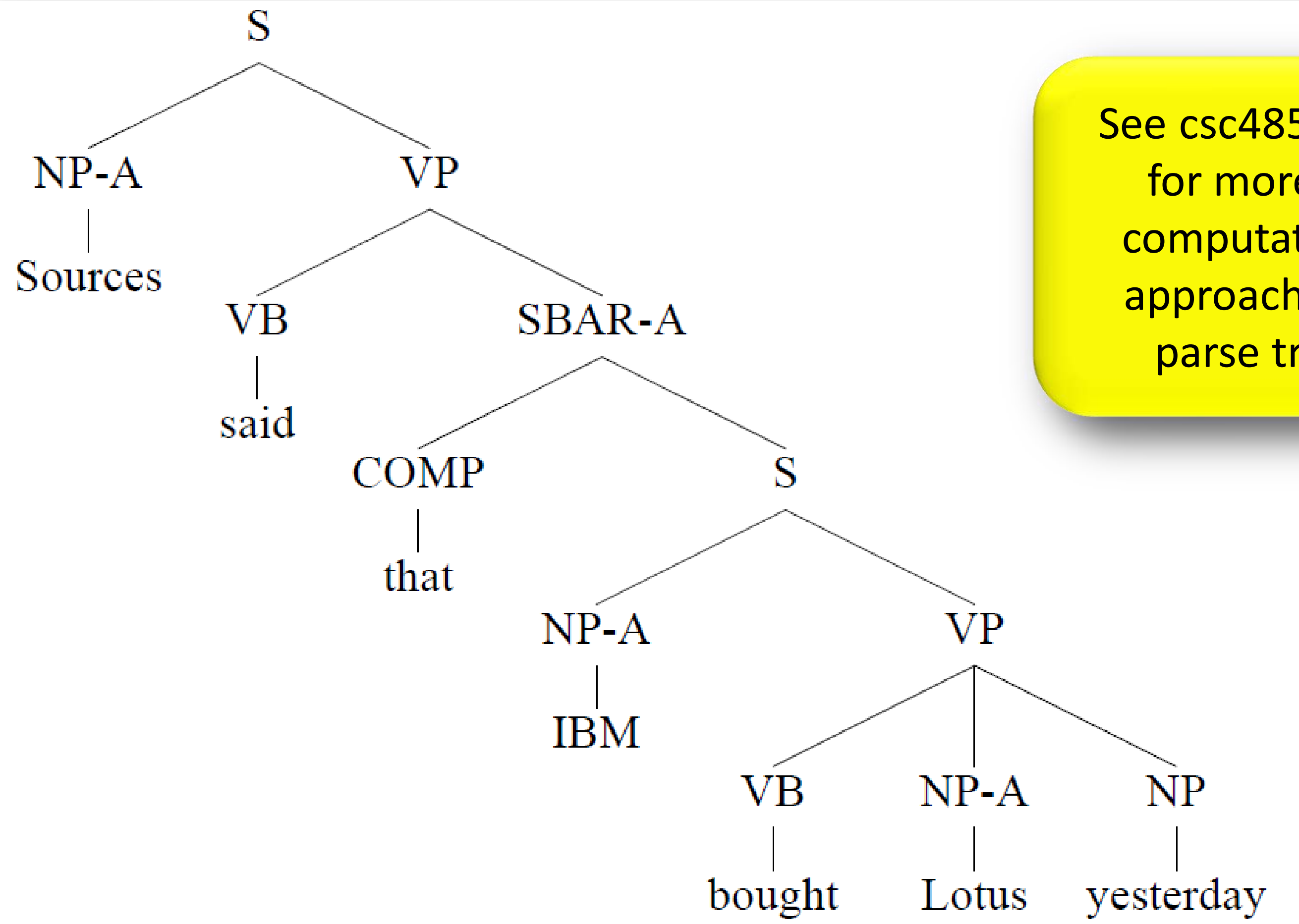
Classical MT: Transfer-based approach

- **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

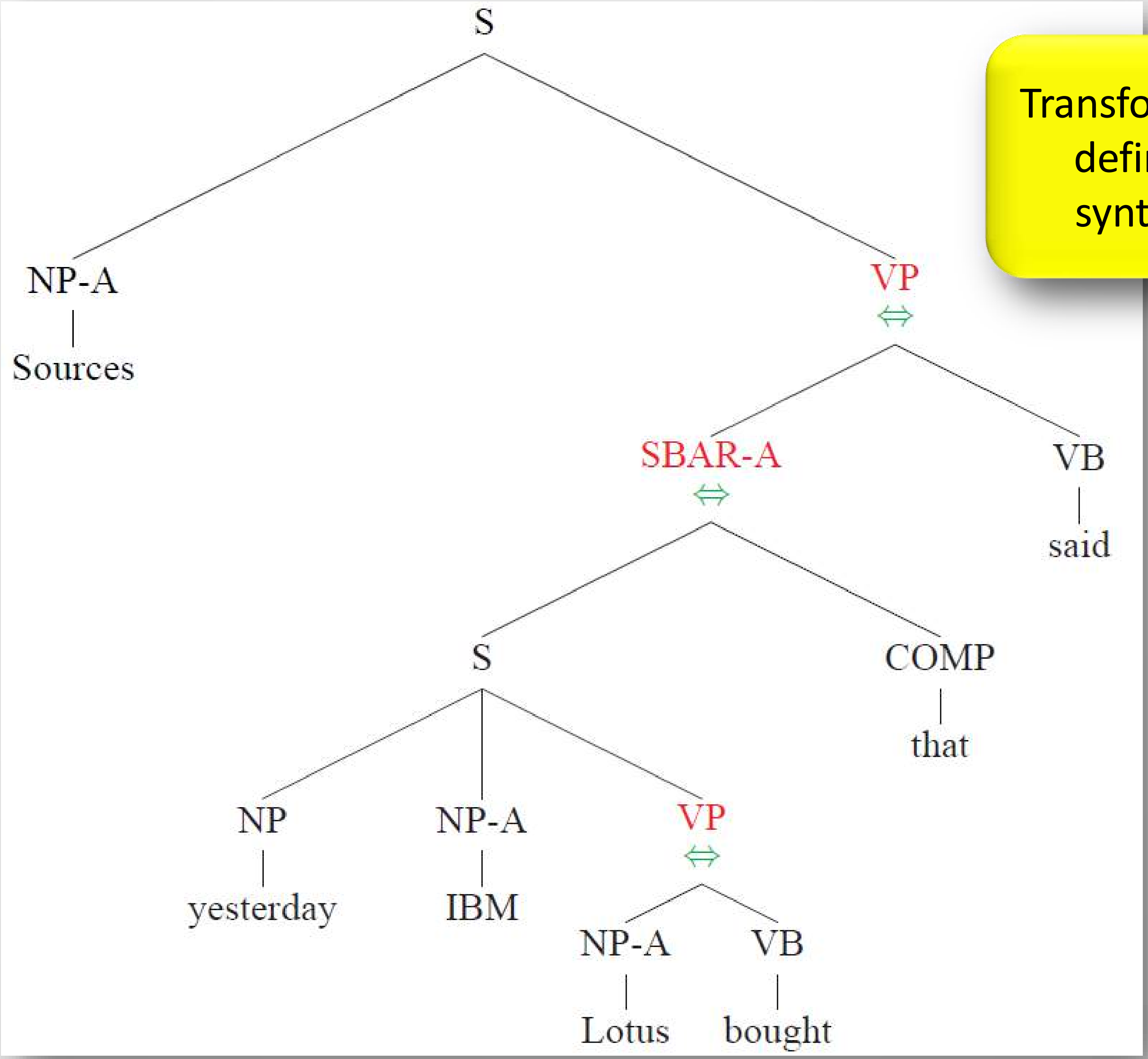


⇒ Japanese: *Sources yesterday IBM Lotus bought that said*

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

From Regina Barzilay at MIT

Example of syntactic transfer



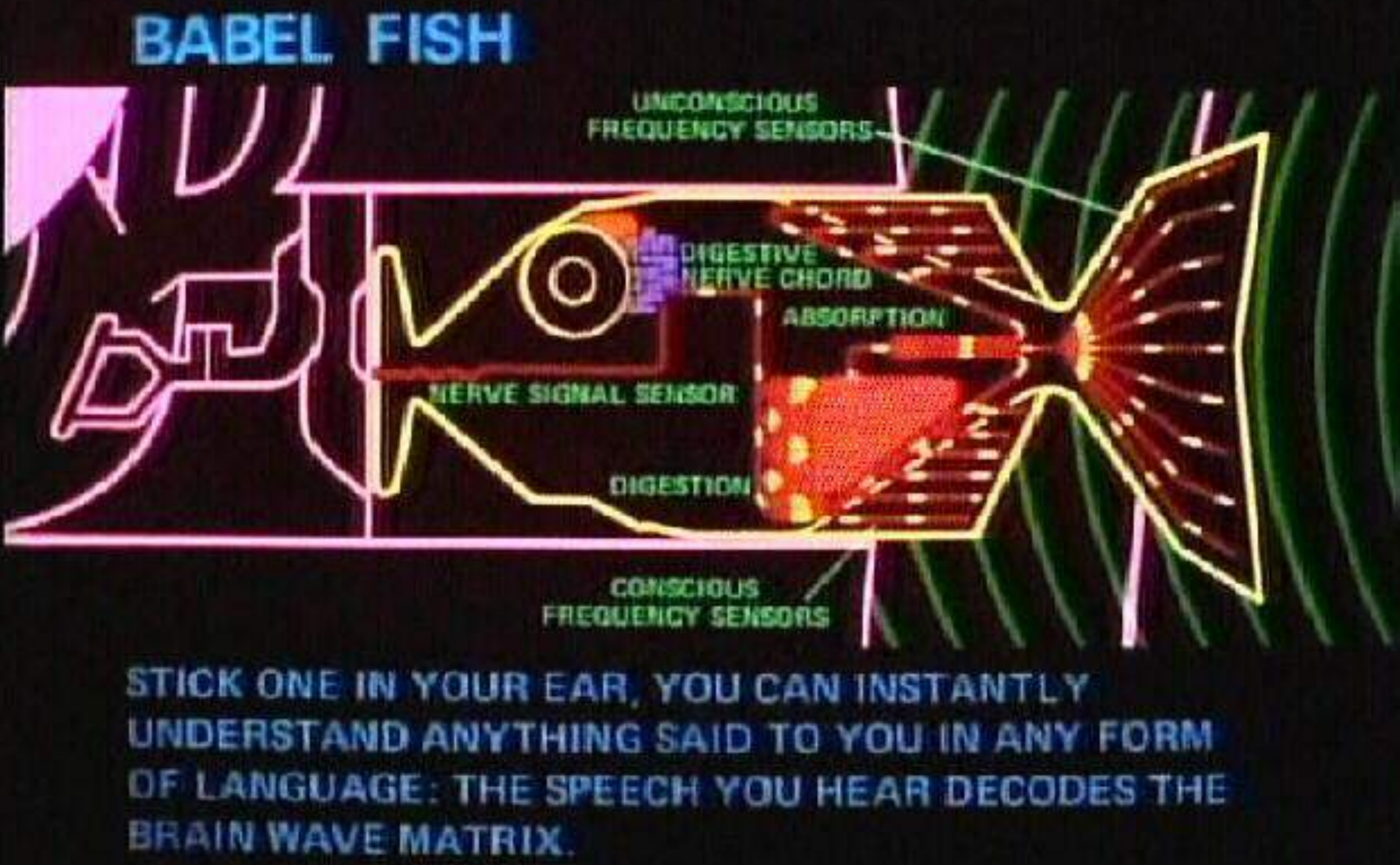
Transformations are defined at the syntactic level

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.

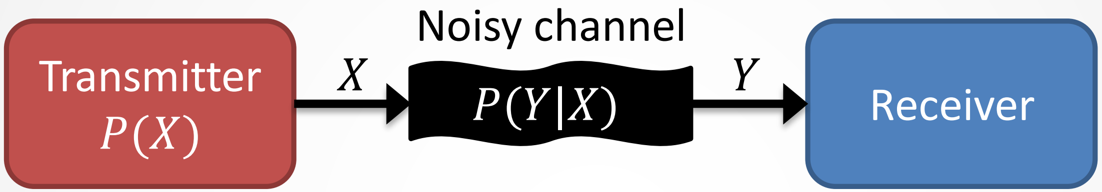
- However, if we want to translate between L languages, we'd need $O(L^2)$ sets of transformation rules.
  - This would involve lots of experts in each language (**$$**).
  - This can be somewhat **mitigated** by abstracting beyond syntax into an **interlingua**: a conceptual space common to **all**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



THE NOISY CHANNEL

The noisy channel

- Messages can get **distorted** when passed through a **noisy** conduit



With great power comes great responsibility → *With great ability comes great accountability*

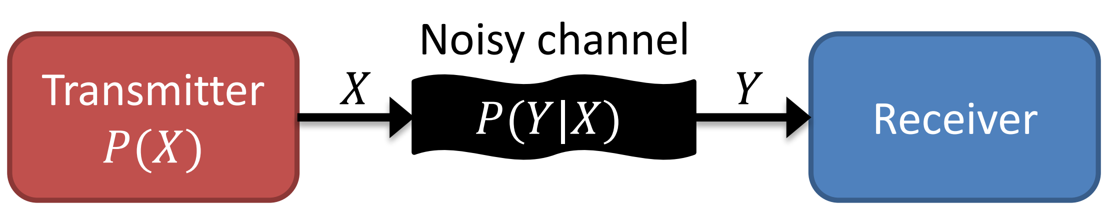
The blue house → *La maison bleue*

Statistical machine translation

- Machine translation seemed to be an intractable problem 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., $$E^* = \underset{E}{\operatorname{argmax}} P(E, F)$$

- Others may be tempted to model this as

$$E^* = \underset{E}{\operatorname{argmax}} P(E|F)P(F)$$

X

This is useless if you're always given F

UNIVERSITY OF TORONTO

How not to use the noisy channel

- Others may be tempted to model this as

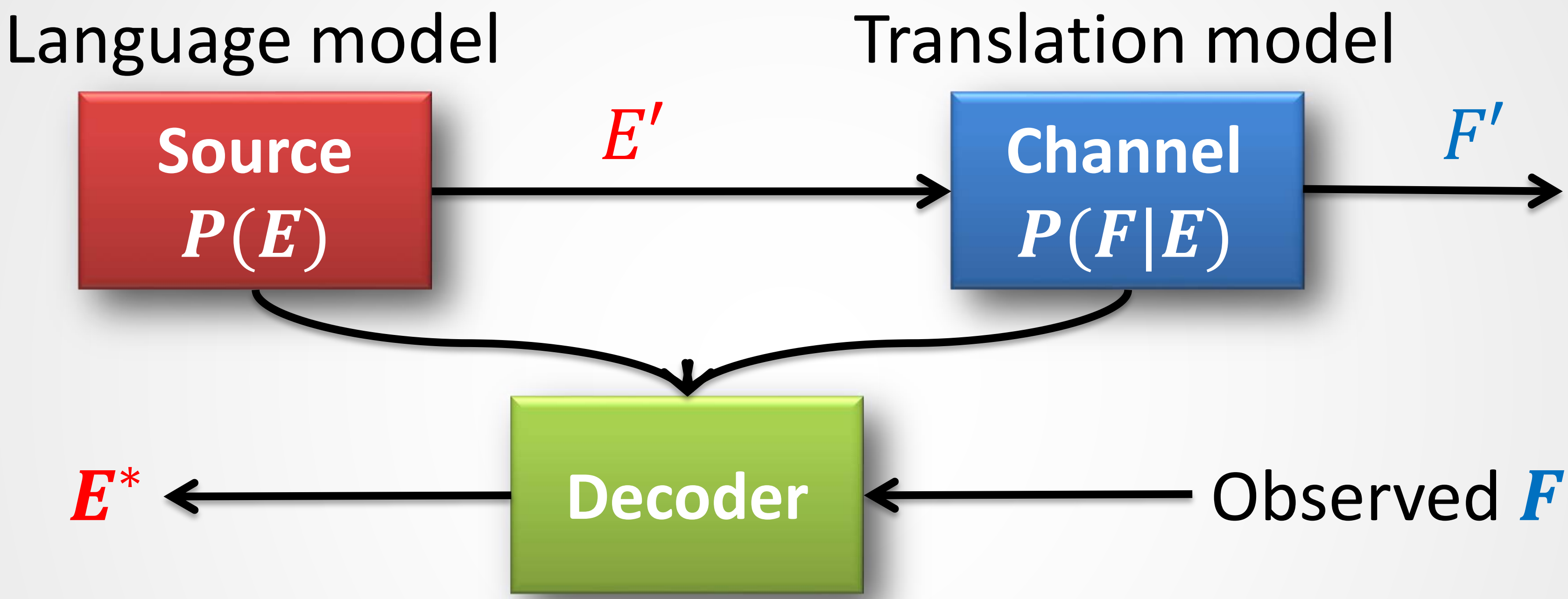
$$E^* = \underset{E}{\operatorname{argmax}}\, P(E|F)\underline{P(F)}$$

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



$$E^* = \underset{E}{\operatorname{argmax}} P(F|E)P(E)$$

How to use the noisy channel

- How does this work?

$$E^* = \underset{E}{\operatorname{argmax}}\, P(F|E)P(E)$$

- $P(E)$ 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*
  →
  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 $P(F|E)$?

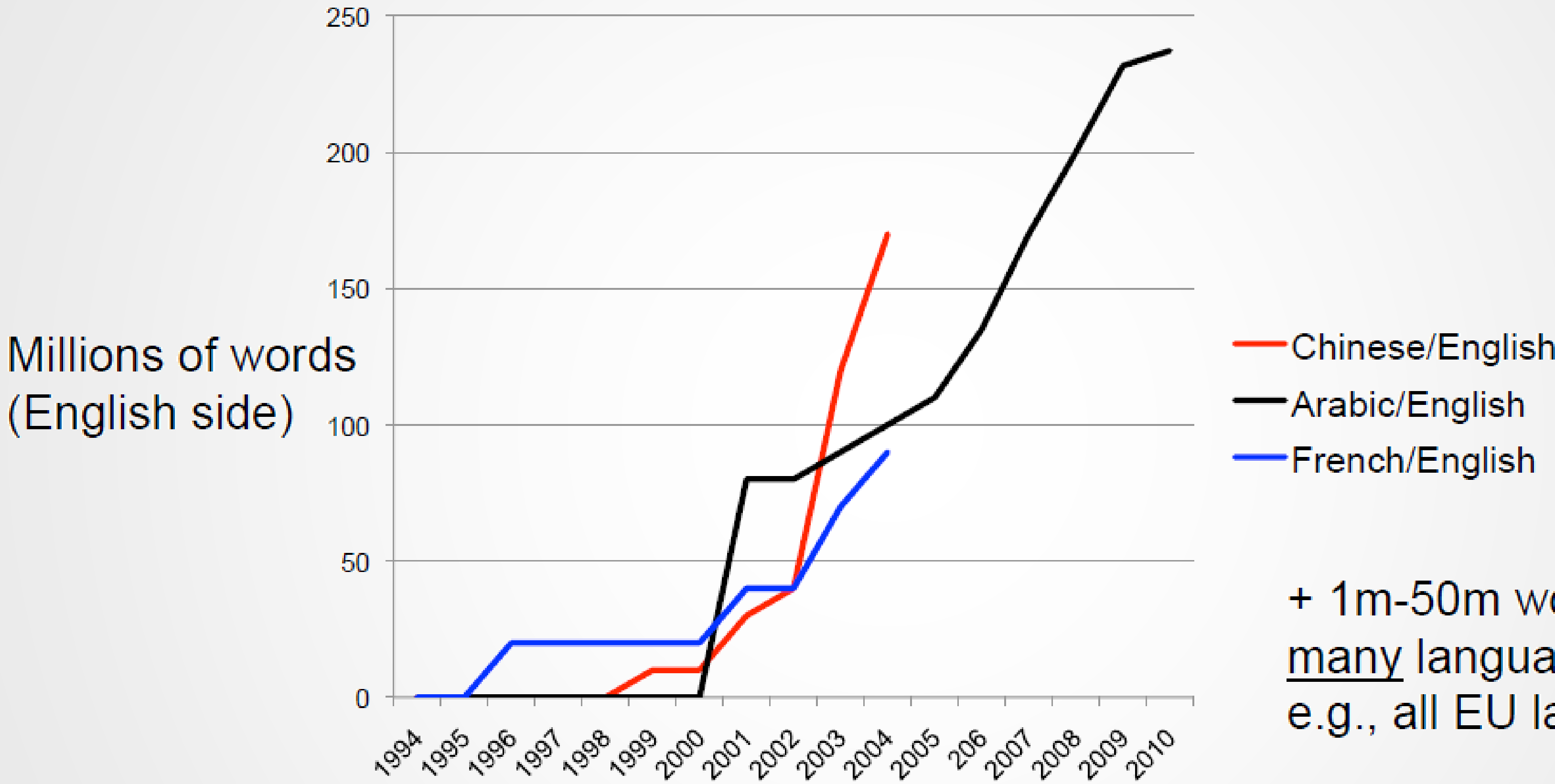
- Solution: collect statistics on vast parallel texts

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

… **<u>citoyen</u>** canadien a le **<u>droit</u>** de vote et est éligible aux élections législatives fédérales ou provinciales …

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

Bilingual data



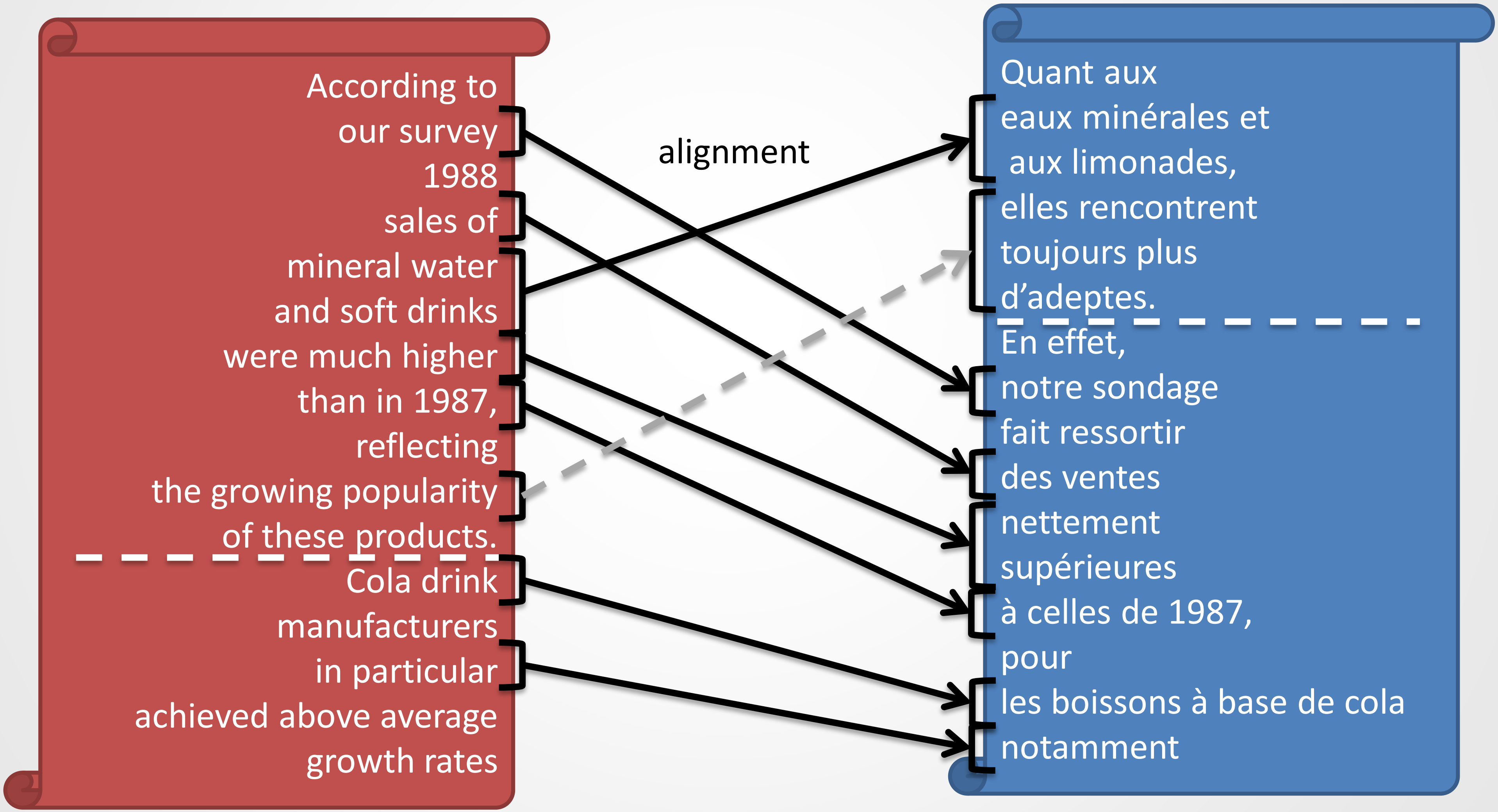
+ 1m-50m words for many language pairs, e.g., all EU languages)

From Chris Manning's course at Stanford

- Data from Linguistic Data Consortium at University of Pennsylvania.

Alignment

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



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.*$_{E1}$ *He had bacon.*$_{E2}$ → *Il était heureux parce qu'il avait du bacon.*$_{F1}$

E_1	F_1
E_2	F_2
E_3	F_3
E_4	F_4
E_5	F_5
E_6	F_6
E_7	F_7
…	

⇒

E_1 E_2	F_1
E_3	F_2
E_4	F_3
E_5	F_4 F_5
E_6	F_6
E_7	F_7
…	

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,
  - $\mathcal{L}_E$ is the # of words in an English sentence,
  - $\mathcal{L}_F$ is the # of words in a French sentence.

- Assume $\mathcal{L}_E$ and $\mathcal{L}_F$ have Gaussian/normal distributions with $\boldsymbol{\mu = c\mathcal{L}_X}$ and $\boldsymbol{\sigma^2 = s^2\mathcal{L}_X}$.
  - **Empirical** constants c and s set 'by hand'.
  - The **penalty**, $Cost(\mathcal{L}_E, \mathcal{L}_F)$, of aligning sentences with different lengths is based on the *divergence* of these Gaussians.

1. Sentence alignment by length

E_1 E_2	F_1
E_3	F_2
E_4	F_3
E_5	F_4 F_5
E_6	F_6

It's a bit more complicated – see paper on course webpage

We can associate costs with different **types** of alignments.

$\boldsymbol{C_{i,j}}$ is the prior cost of aligning i sentences to j sentences.

$$
\begin{aligned}
Cost = {} & Cost(\mathcal{L}_{E_1} + \mathcal{L}_{E_2}, \mathcal{L}_{F_1}) + C_{2,1} + \\
& Cost(\mathcal{L}_{E_3}, \mathcal{L}_{F_2}) + C_{1,1} + \\
& Cost(\mathcal{L}_{E_4}, \mathcal{L}_{F_3}) + C_{1,1} + \\
& Cost(\mathcal{L}_{E_5}, \mathcal{L}_{F_4} + \mathcal{L}_{F_5}) + C_{1,2} + \\
& Cost(\mathcal{L}_{E_6}, \mathcal{L}_{F_6}) + C_{1,1}
\end{aligned}
$$

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., $Count(s, h, i)$ is a **trigraph** model

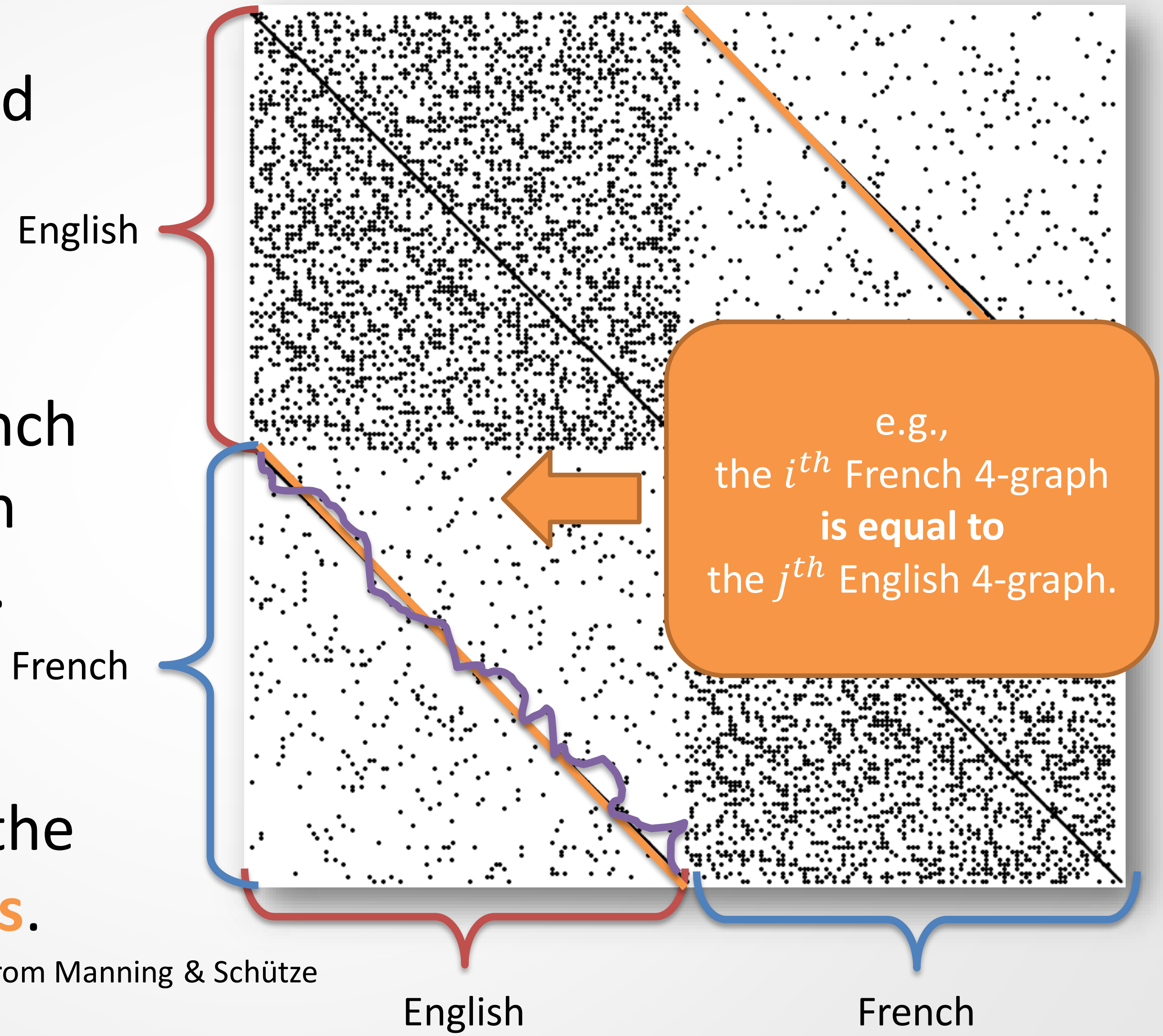
- Church (1993) tracks all **4-graphs** which are identical across two texts.
  - He calls this a 'signal-based' approximation to cognate identification.

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**.

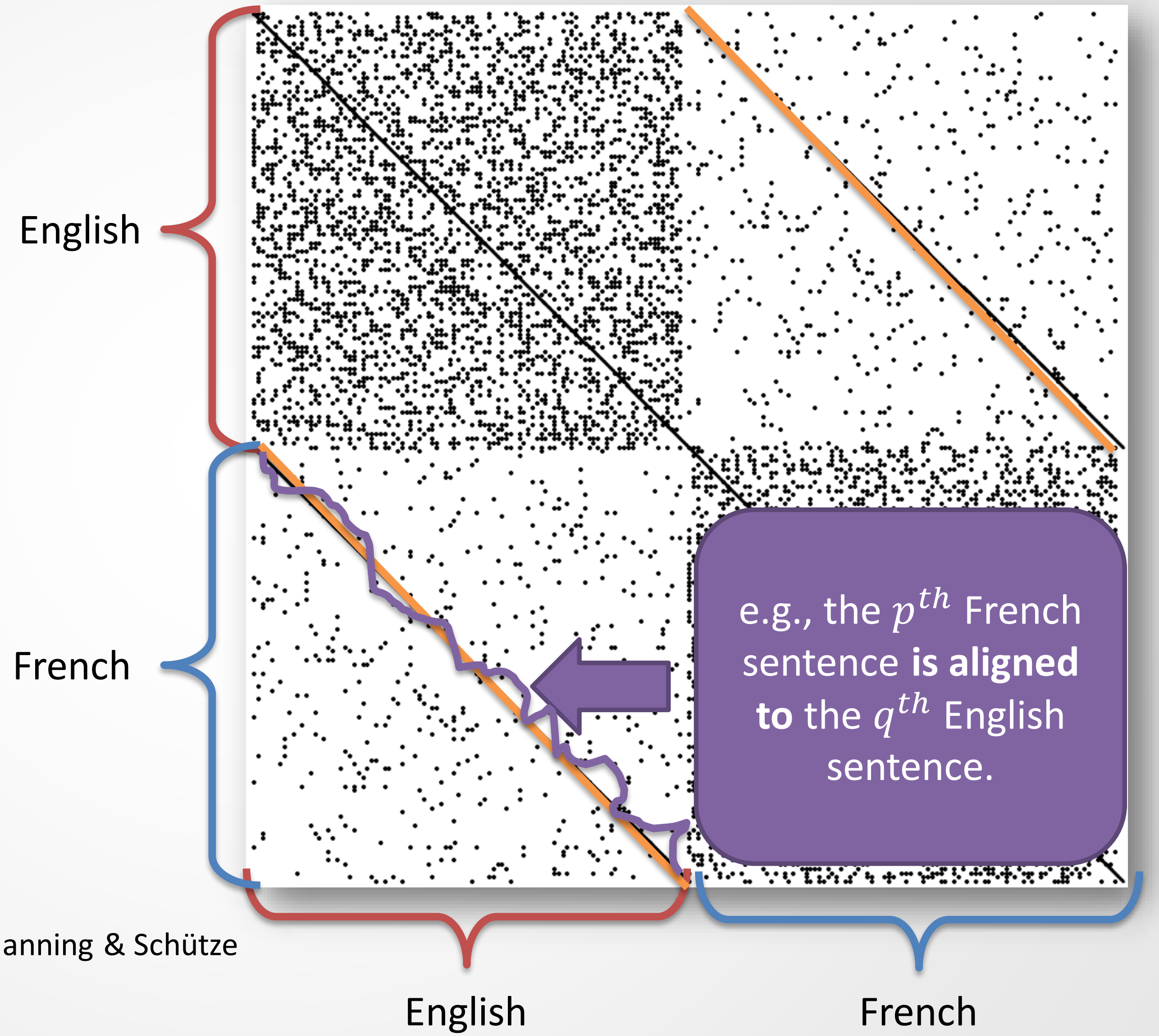


From Manning & Schütze

UNIVERSITY OF TORONTO

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

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

$$LCSR(A, B) = \frac{length(LCS(A, B))}{\max(length(A), length(B))}$$

  - e.g.,

$$LCSR(\mathbf{government}, \mathbf{gouvernement}) = \frac{10}{12}$$

'LCS Ratio'

2b. Melamed's method

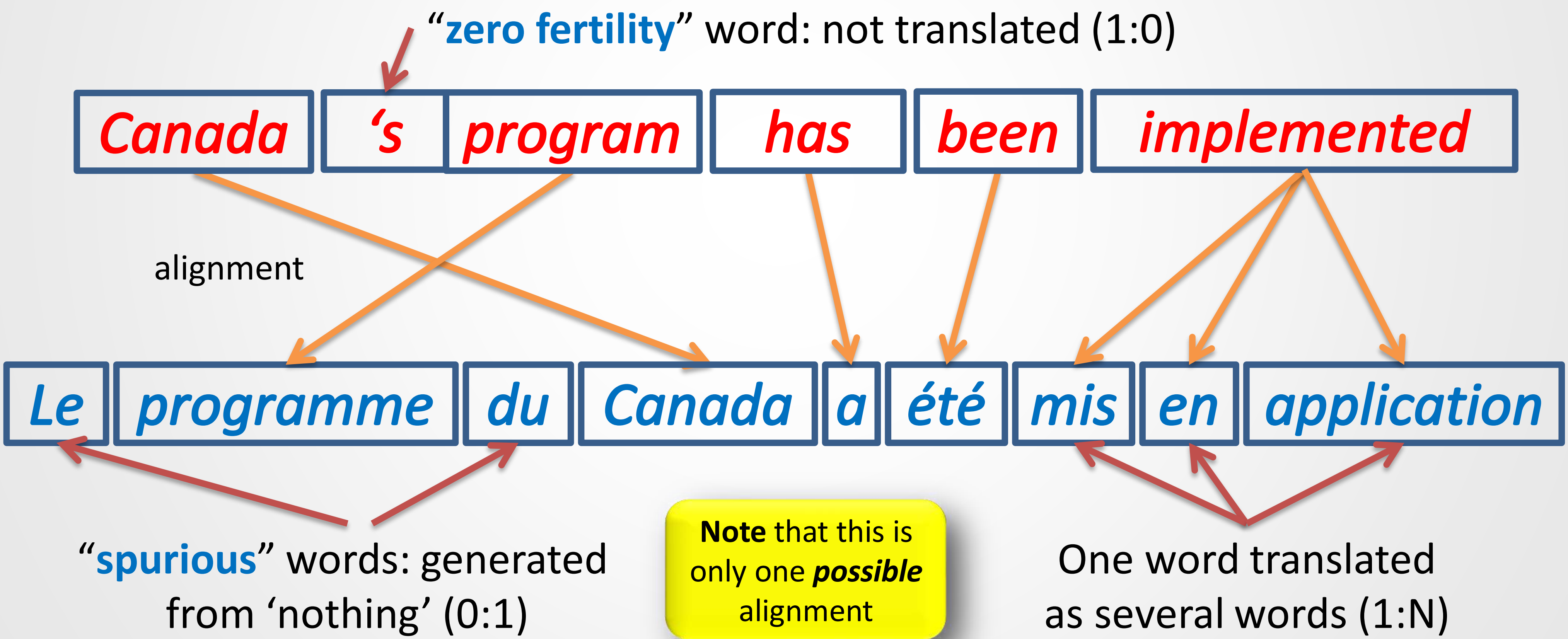
- Excludes **stop words** from both languages.
  (e.g., *the*, *a*, *le*, *un*)
- Melamed empirically declared that cognates occur when $\boldsymbol{LCSR \geq 0.58}$ (i.e., there's a lot of overlap in those words).
  - $\therefore$ 25% of words in Canadian Hansard are cognates.
- As with Church, construct a "bitext" **graph**.
  - Put a point at position $(i,j) \equiv LCSR(i,j) \geq 0.58$.
  - Find a near-diagonal alignment, as before.

From sentences to words

- We've computed the **sentence** alignments.
- What about **word** alignments?

Word alignment

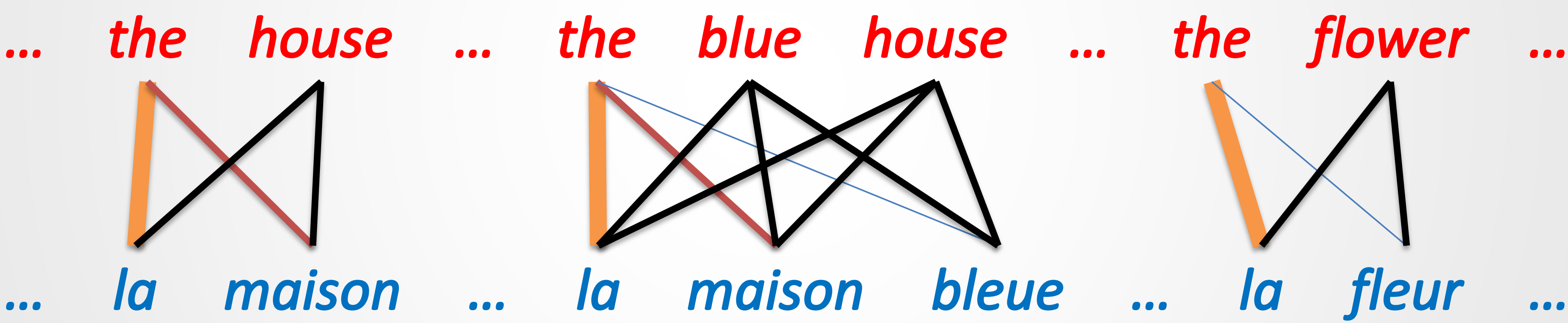
- **Word alignments** can be 1:1, N:1, 1:N, 0:1,1:0,… E.g.,



UNIVERSITY OF TORONTO

Intuition of statistical MT

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



$P(la|the)$ should be **higher** than $P(fleur|the)$, $P(bleue|the)$, and even $P(maison|the)$

Note: we consider **all possible** word alignments….

Reading

- **Entirely optional**: Vogel, S., Ney, H., and Tillman, C. (1996). *HMM-based Word Alignment in Statistical Translation*. In: Proceedings of the 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen.

- (optional) Gale & Church "*A Program for Aligning Sentences in Bilingual Corpora*" (on course website)

- **Useful reading on IBM Model-1**: Section 25.5 of the 2nd edition of the Jurafsky & Martin text.
  - 1st edition available at Robarts library.

- **Other**: Manning & Schütze Sections 13.0, 13.1.2 (Gale&Church), 13.1.3 (Church), 13.2, 13.3, 14.2.2

UNIVERSITY OF TORONTO