

# HMMs Summary

- Important ideas to know:
  - The definition of an HMM (e.g., its parameters).
  - The purpose of the **Forward algorithm**.
    - How to compute  $\alpha_i(t)$  and  $\beta_i(t)$
  - The purpose of the **Viterbi algorithm**.
    - How to compute  $\delta_i(t)$  and  $\psi_i(t)$ .
  - The purpose of the **Baum-Welch algorithm**.
    - Some understanding of EM.
    - Some understanding of the equations.

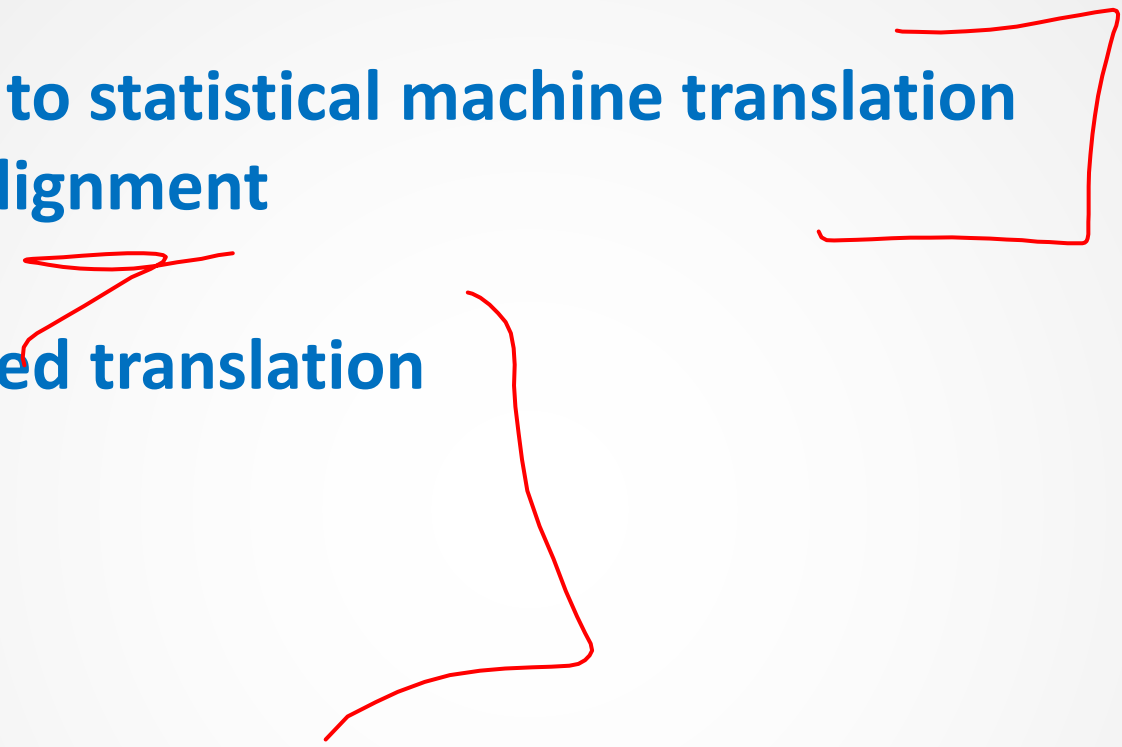


# statistical machine translation

## ***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
- 



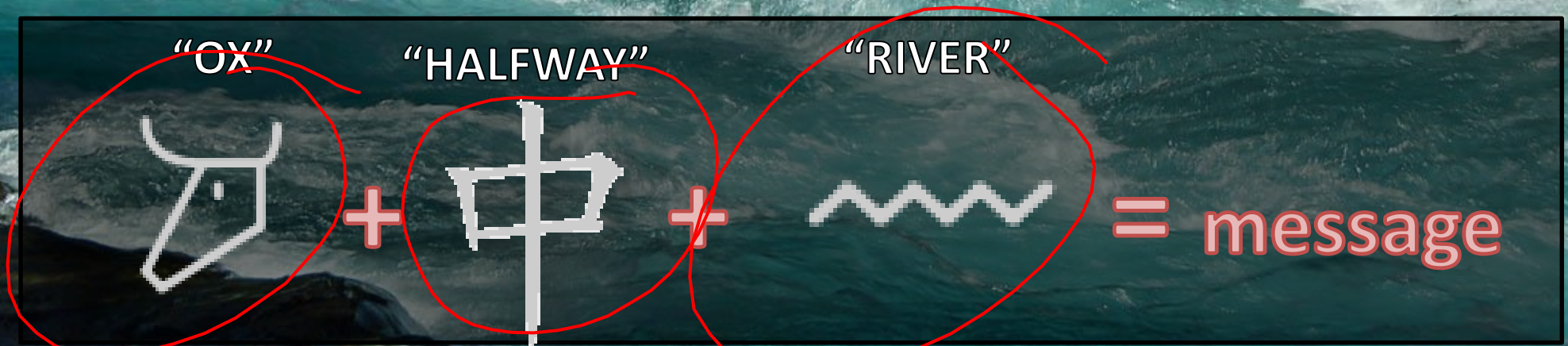
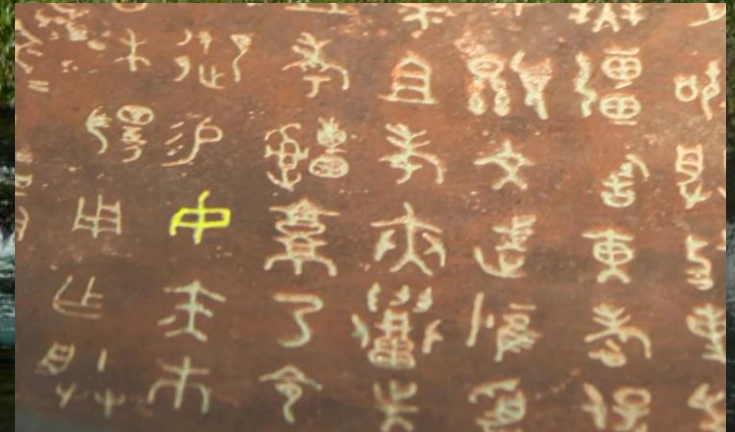
# THE ABSTRACTIONS OF BEASTS



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



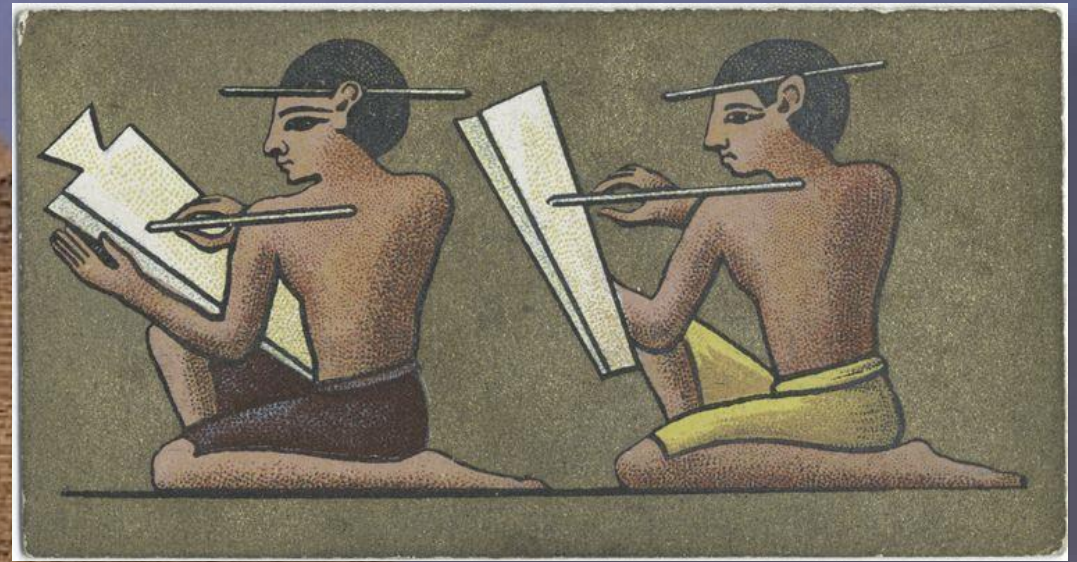
Imagine your ancestor  
wanted to leave the  
message *"there are ox  
halfway up the river"*



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)



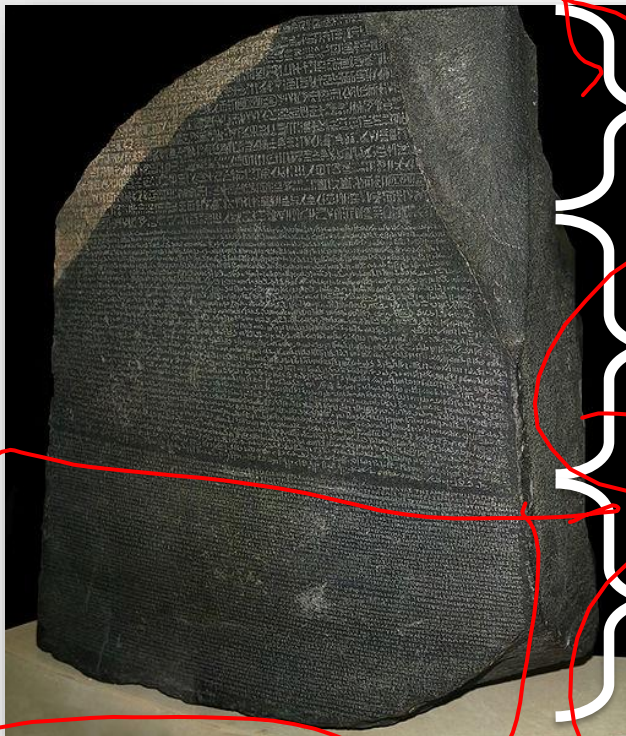
### • Demotic (c. 650 BCE)

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



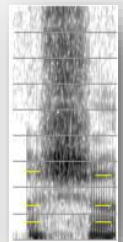
Ancient  
Egyptian  
hieroglyphs

Egyptian  
Demotic

Ancient  
Greek

# 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




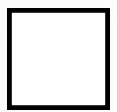










- Some writing systems are a mix of these qualities:

- 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?***

# Evolution of the Alphabet

Proto-Sinaitic  
c. 1750 BCE

𐤀 𐤁 𐤂 𐤃 𐤄 𐤅 𐤆 𐤇 𐤈 𐤉 𐤊 𐤋 𐤌 𐤍 𐤎 𐤏 𐤐 𐤑 𐤒 𐤓 𐤔 𐤕

Phoenician  
c. 1000 BCE

𐤀 𐤁 𐤂 𐤃 𐤄 𐤅 𐤆 𐤇 𐤈 𐤉 𐤊 𐤋 𐤌 𐤍 𐤎 𐤏 𐤐 𐤑 𐤒 𐤓 𐤔 𐤕

Archaic Greek  
c. 750 BCE

Α Β Γ Δ Ε Ζ Η Θ Ι Κ Λ Μ Ν Ξ Ο Π Ρ Σ Τ Υ Φ Χ Ψ

Archaic Latin  
c. 500 BCE

A B C D E F G H I K L M N O P Q R S T V X

Roman  
c. 1 CE

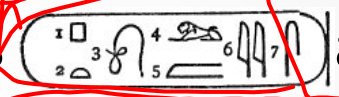

A B C D E F G H I K L M N O P Q R S T V X Y Z

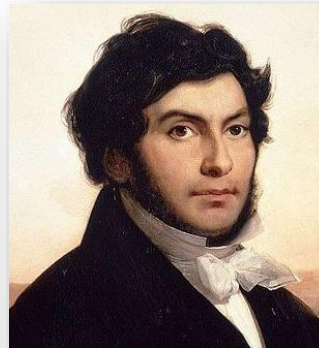
Modern Latin  
Script

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

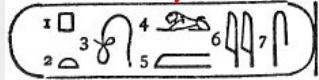



# Deciphering Rosetta


- During 1822–1824, **Jean-François Champollion** worked on the Rosetta stone. He noticed:
  - The circled Egyptian symbols  appeared in roughly the same positions as the word 'Ptolemy' in the Greek.
  - The number of Egyptian hieroglyph tokens were **much larger** than the number of Greek words → Egyptian seemed to have been partially phonographic.
  - 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 (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 lance pas ce bagel!*



*Don't throw that bagel!*



# Direct translation

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

<i>¿</i>	<i>Dónde</i>	<i>está</i>	<i>la</i>	<i>biblioteca</i>	<i>?</i>
	<i>Where</i>	<i>is</i>	<i>the</i>	<i>library</i>	<i>?</i>
	<i>Où</i>	<i>est</i>	<i>la</i>	<i>bibliothèque</i>	<i>?</i>

<i>Mi</i>	<i>nombre</i>	<i>es</i>	<i>T-bone</i>
<i>My</i>	<i>name</i>	<i>is</i>	<i>T-bone</i>
<i>Mon</i>	<i>nom</i>	<i>est</i>	<i>T-bone</i>



# 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 cueva flotando*  
(*the bottle entered to the cave floating*)



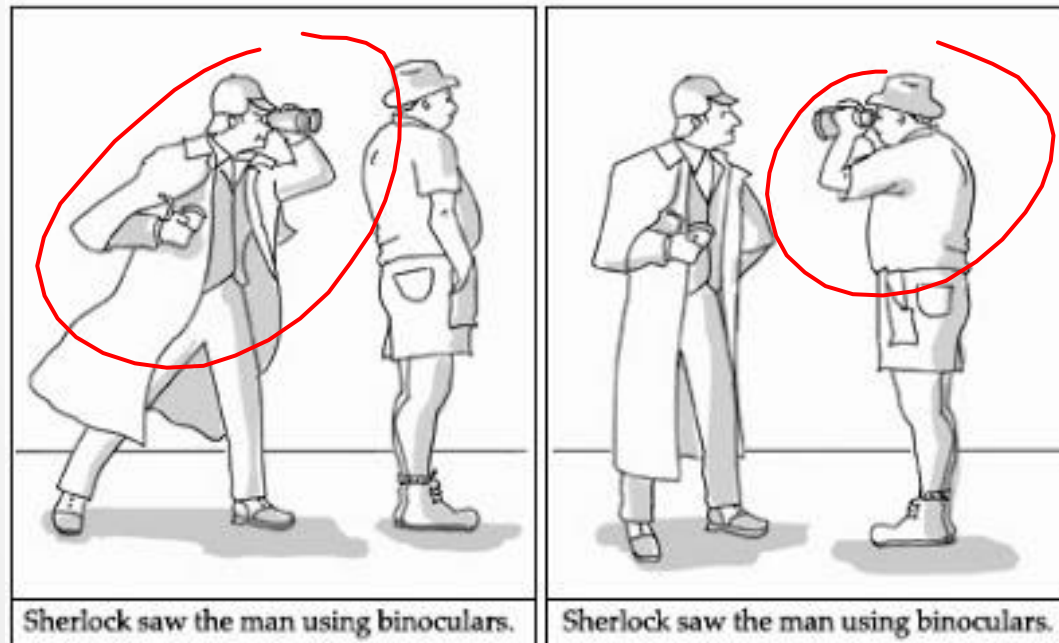
*The bottle floated into the cave*

- 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*



# Challenge 4: syntactic ambiguity

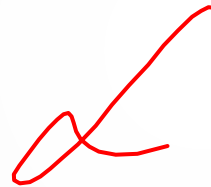
- **Syntactic ambiguity** in the source makes it difficult to produce a single sentence in the target language.
  - *E.g.,*

*Rick hit the Morty with the stick*



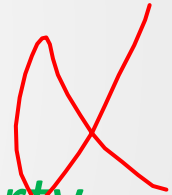
*Rick golpeó el Morty  
con el palo*

(the stick was used)



*Rick golpeó el Morty  
que tenia el palo*

(the Morty had the stick)



# Challenge 5: idiosyncracies

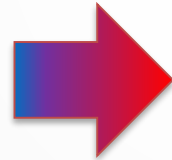
- Languages have their own idioms, and “feel”.
  - E.g.,

*We have to burn the  
midnight oil*



*Il faut travailler tard* ✓  
*Il faut brûler l'huile* X  
*de minuit*

*Estie de sacramouille*



*Host of the sacrament* X

*By golly!* ✓

*L'eau dans la cave*



*Water in the basement* X

*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

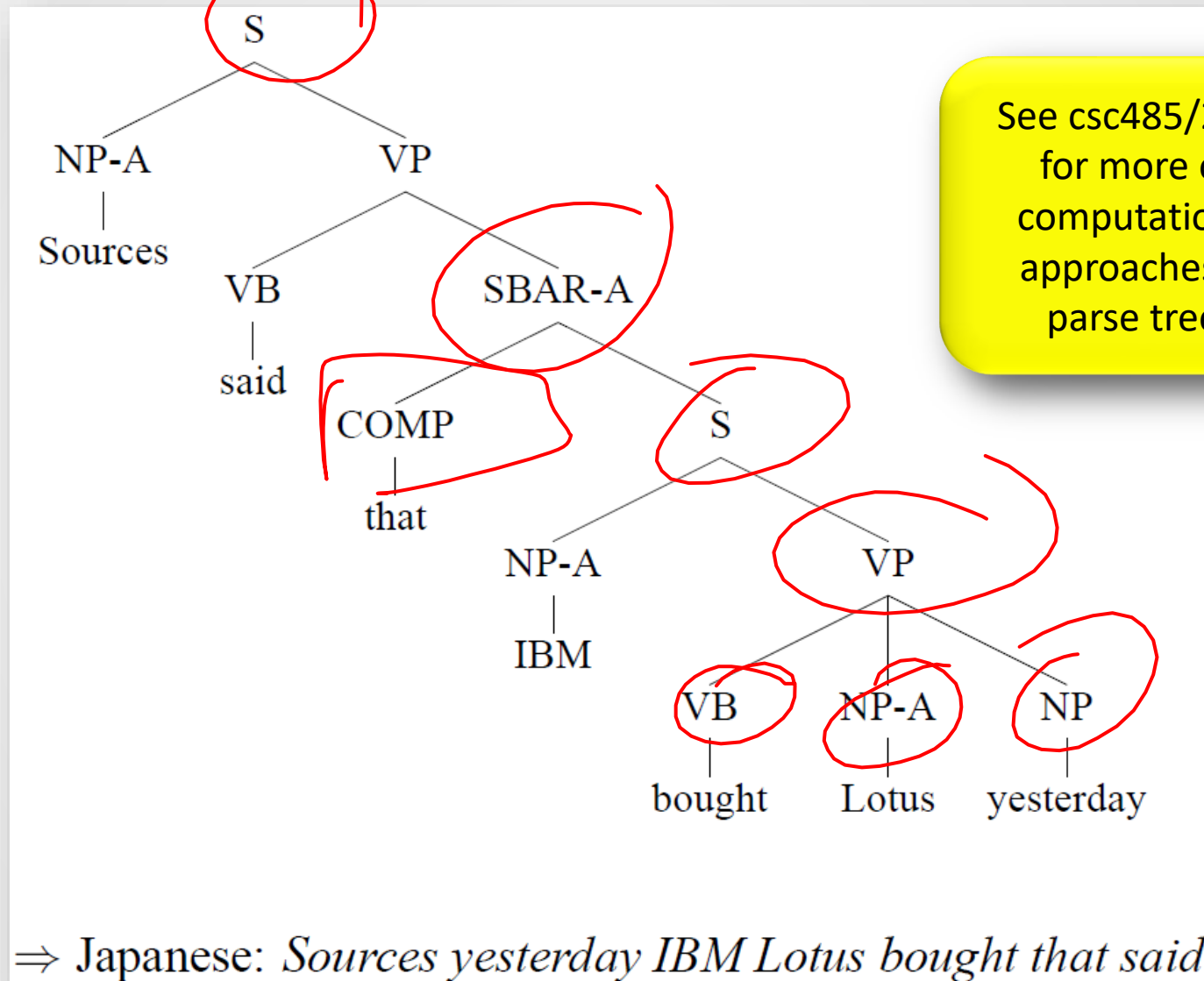
- This approach causes some problems, e.g.,  
SVO
- It's difficult/impossible to capture **long-range** re-orderings:
  - **English:** Sources said that IBM bought Lotus yesterday
  - **Japanese:** ~Sources yesterday IBM Lotus bought that said  
SOV
- It's difficult to disambiguate parts-of-speech:
  - **English:** They said that I punched that Morty
  - **French:** Ils ont dit que j'ai frappé ce 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.

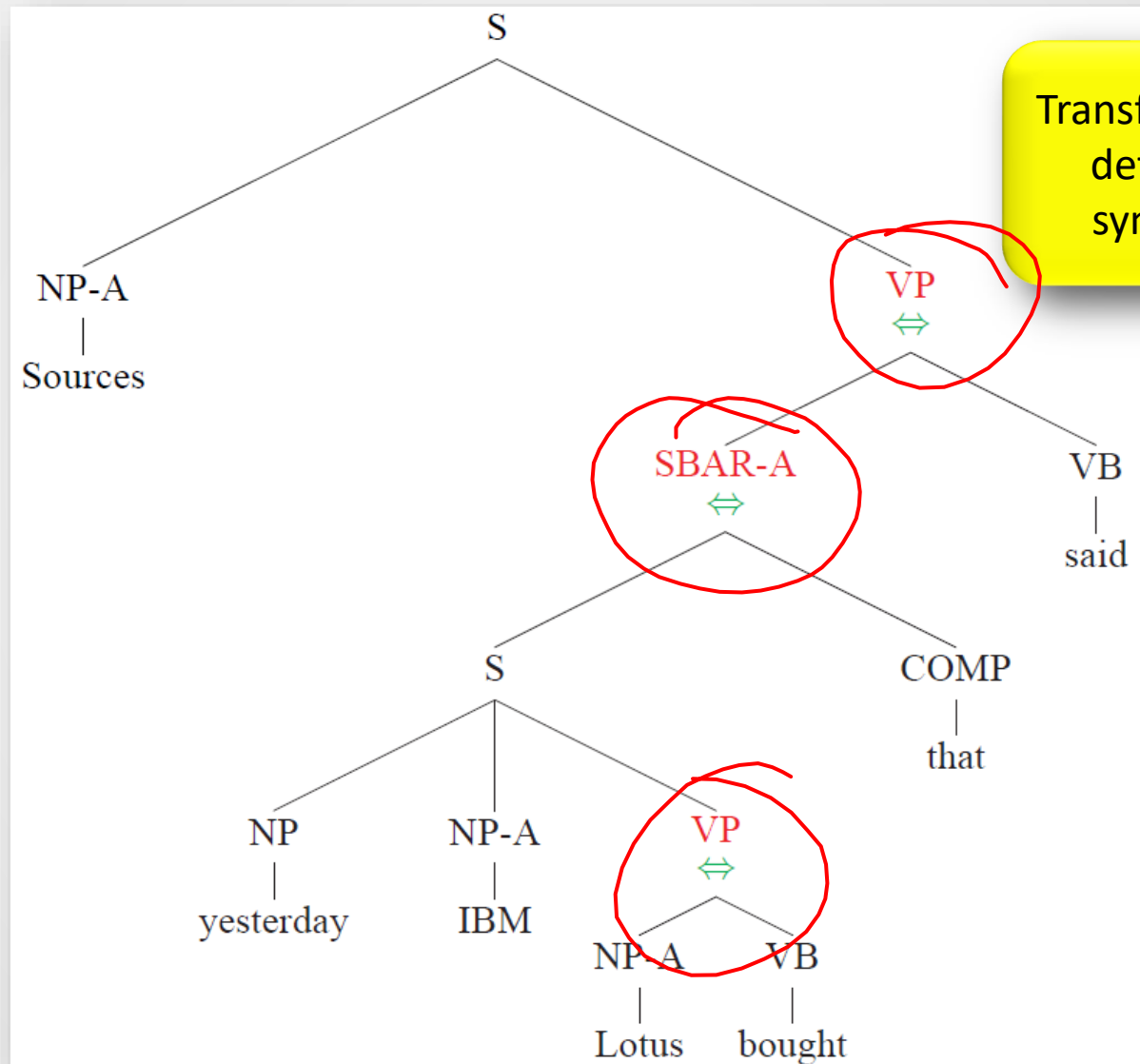


# Example of syntactic transfer



From Regina Barzilay at MIT

# Example of syntactic transfer



Transformations are defined at the syntactic level

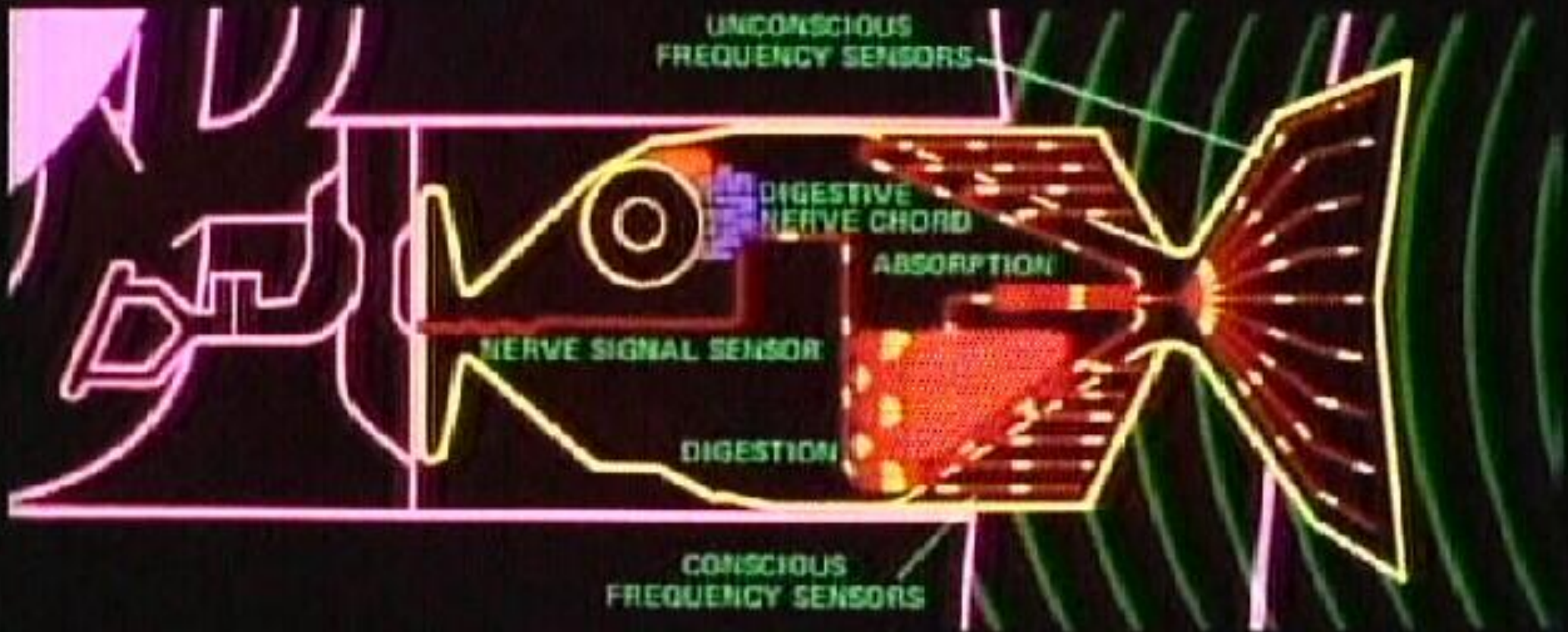
From Regina Barzilay at MIT

# 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.



# BABEL FISH



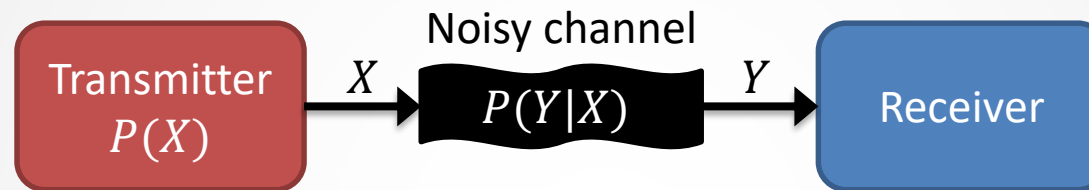
STICK ONE IN YOUR EAR, YOU CAN INSTANTLY UNDERSTAND ANYTHING SAID TO YOU IN ANY FORM OF LANGUAGE: THE SPEECH YOU HEAR DECODES THE BRAIN WAVE MATRIX.

## THE NOISY CHANNEL

# The noisy channel



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

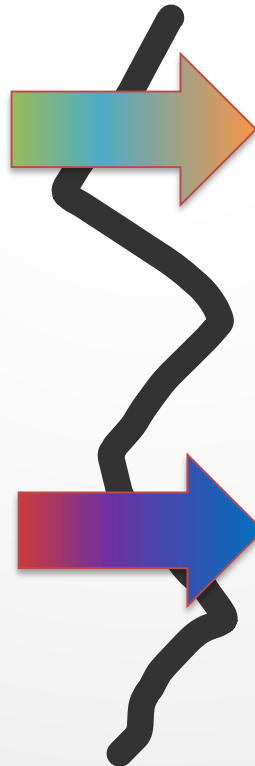


*With great power comes  
great responsibility*

*With great ability comes  
great accountability*

*The blue house*

*La maison bleue*



# Statistical machine translation

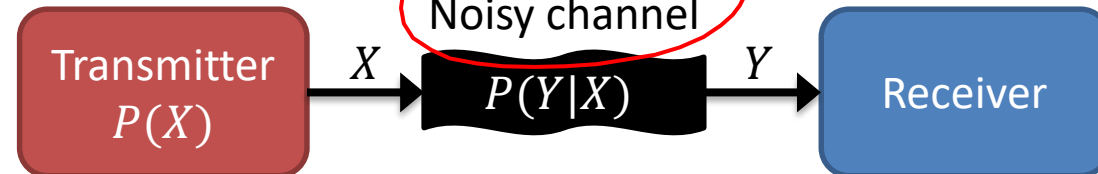
- Machine translation seemed to be an intractable problem until a change in perspective...



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

- Others may be tempted to model this as

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



This is useless if you're  
always given  $F$





# How not to use the noisy channel

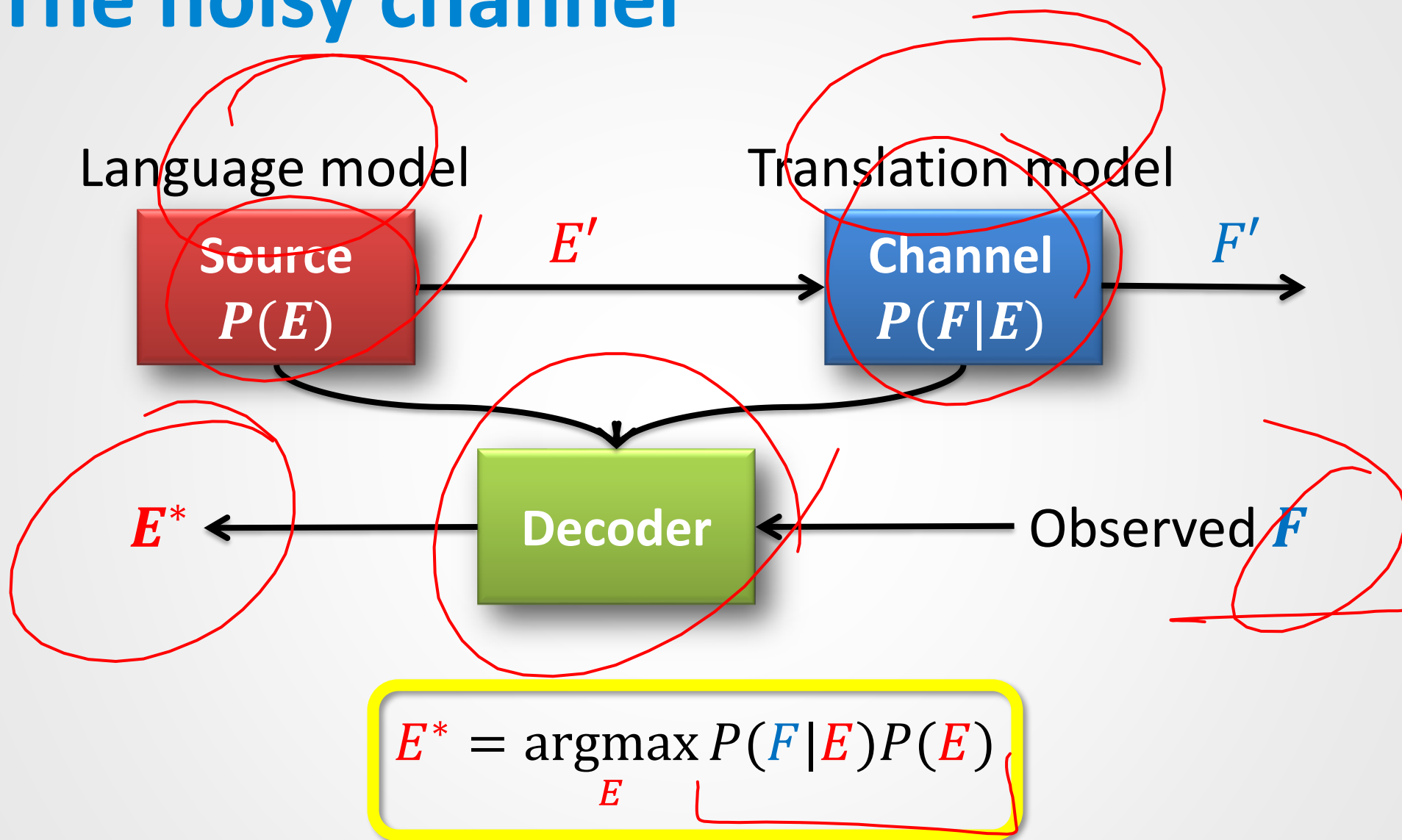
- Others may be tempted to model this as

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

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



# 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*

→

*What hunger have I*

*Hungry I am so*

*I am so hungry*

*Have I that hunger*

...

$$P(S|E) = 1.4E^{-5}$$

$$P(S|E) = 1.0E^{-6}$$

$$P(S|E) = 1.0E^{-6}$$

$$P(S|E) = 2.0E^{-5}$$





# How to use the noisy channel

- ... and with the English language model

- *Que hambre tengo yo*

→

*What hunger have I*

*Hungry I am so*

*I am so hungry*

*Have I that hunger*

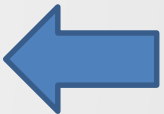
...

$$P(S|E)P(E) = 1.4E^{-5} \times 1.0E^{-6}$$

$$P(S|E)P(E) = 1.0E^{-6} \times 1.4E^{-6}$$

$$P(S|E)P(E) = 1.0E^{-6} \times 1.0E^{-4}$$

$$P(S|E)P(E) = 2.0E^{-5} \times 9.8E^{-7}$$



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

$P(F|E)$   
 $P(E)$

- Solution: collect statistics on vast parallel texts

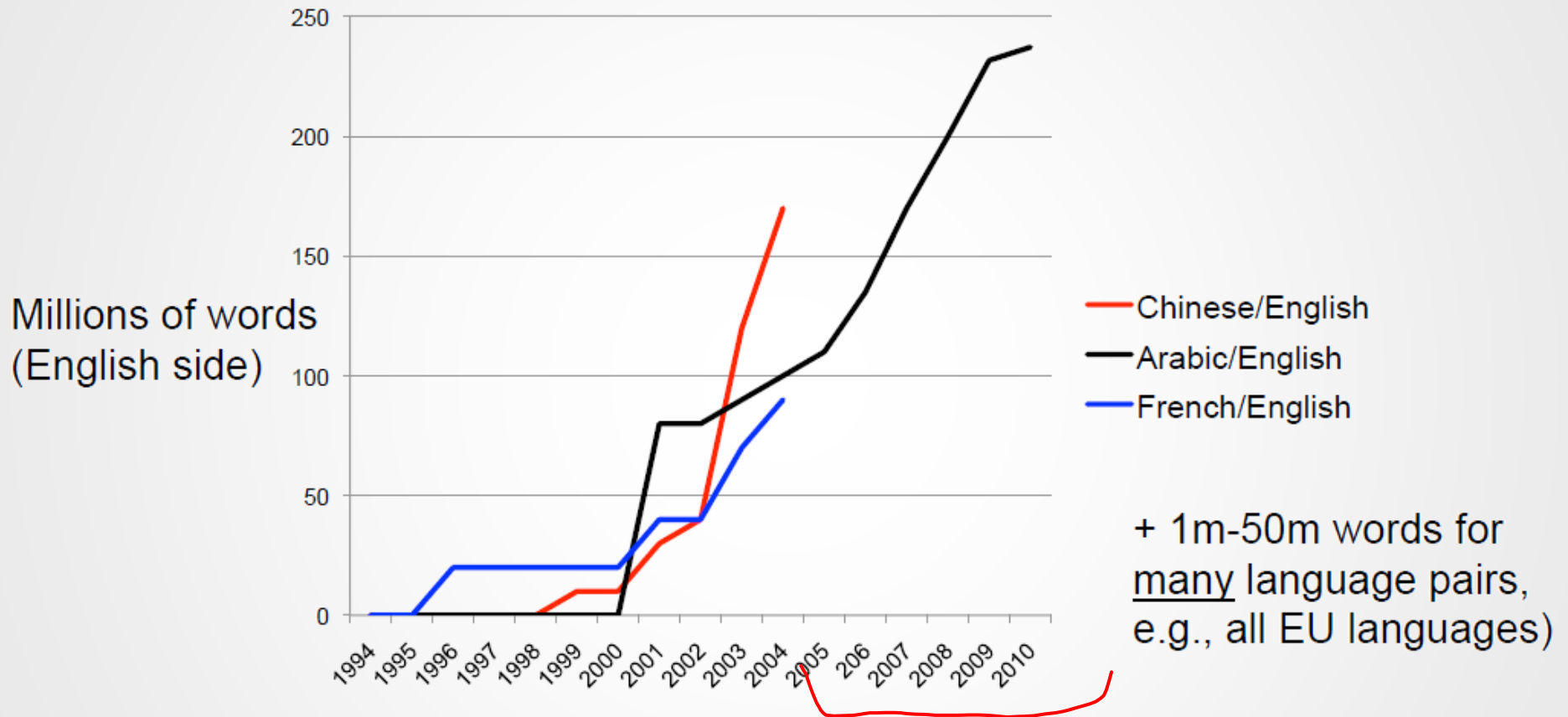
... citizen of  
Canada has the  
right to vote in  
an election of  
members of the  
House of  
Commons or of a  
legislative  
assembly and to  
be qualified for  
membership ...



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

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

# Bilingual data

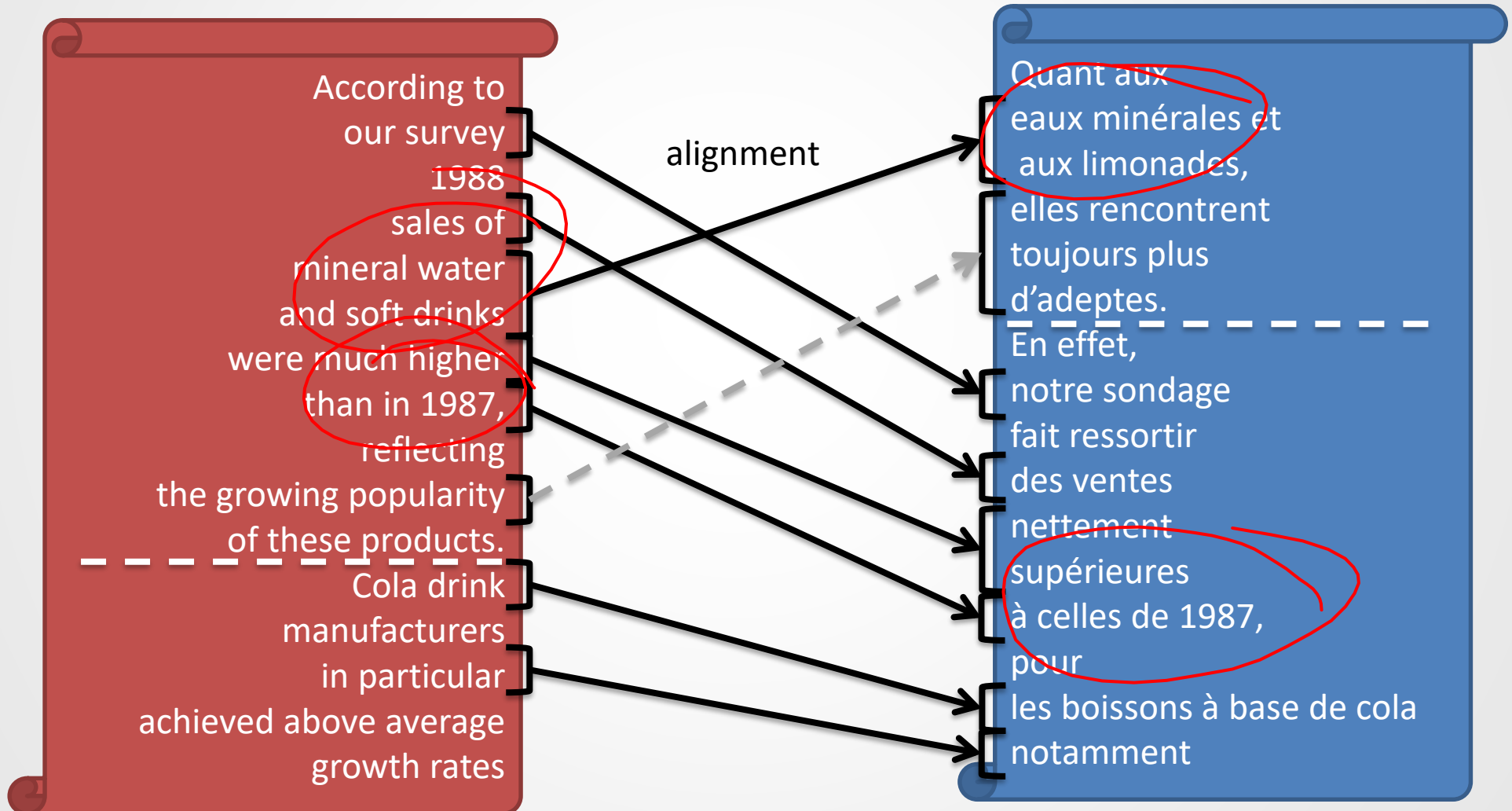


From Chris Manning's course at Stanford

- 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



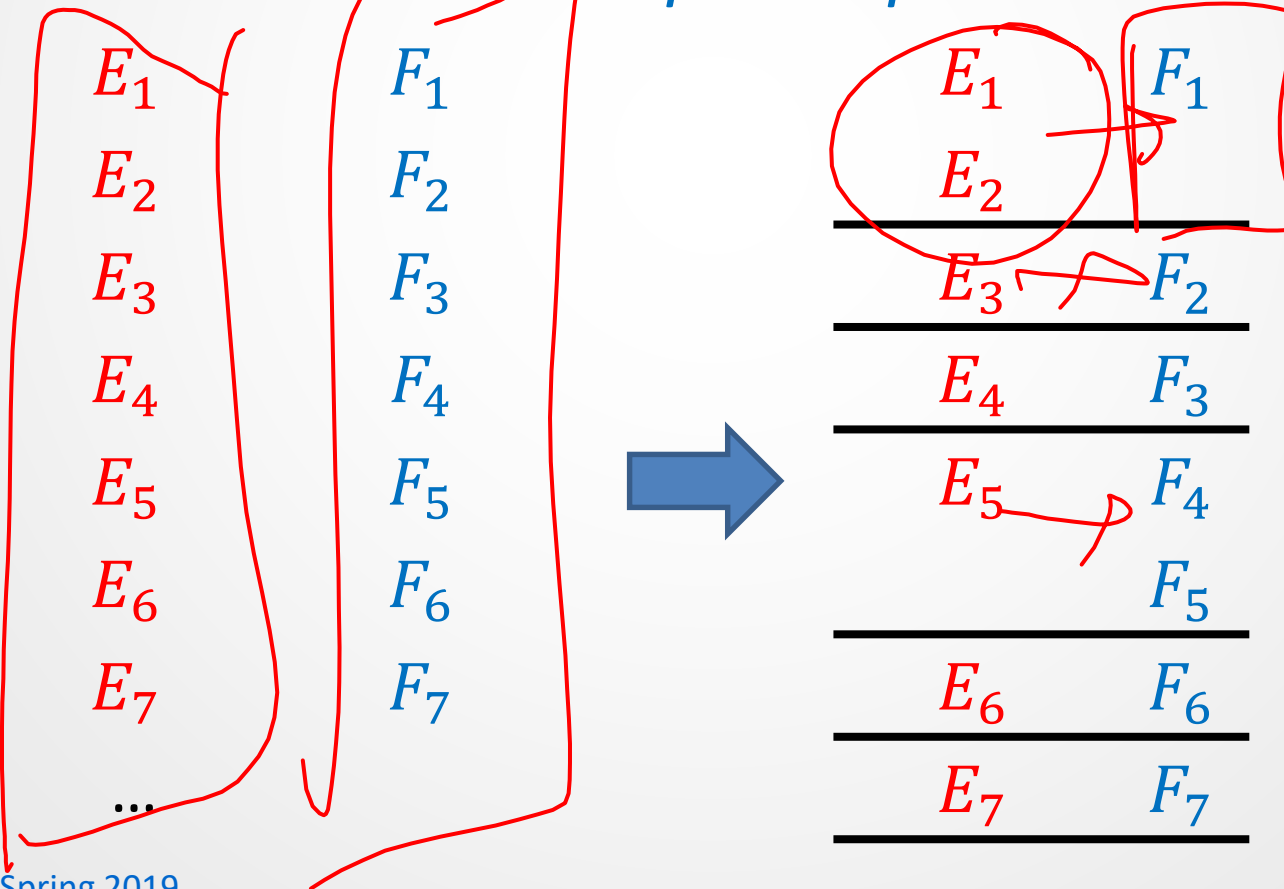
Quant aux  
eaux minérales et  
aux limonades,  
elles rencontrent  
toujours plus  
d'adeptes.  
En effet,  
notre sondage  
fait ressortir  
des ventes  
nettement  
supérieures  
à celles de 1987,  
pour  
les boissons à base de cola  
notamment

From Manning & Schütze

# Sentence alignment

- Sentences can also be **unaligned** across translations.

• E.g., *He was happy.*<sub>E1</sub> *He had bacon.*<sub>E2</sub> → *Il était heureux parce qu'il avait du bacon.*<sub>F1</sub>



# 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”.

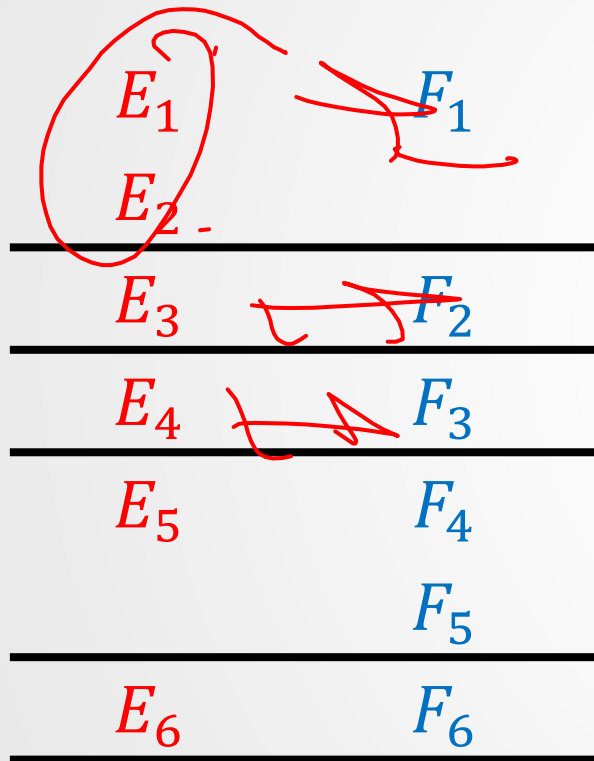
# 1. Sentence alignment by length

(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  $\mu = c\mathcal{L}_X$  and  $\sigma^2 = s^2\mathcal{L}_X$ .
  - **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



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

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

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

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

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.

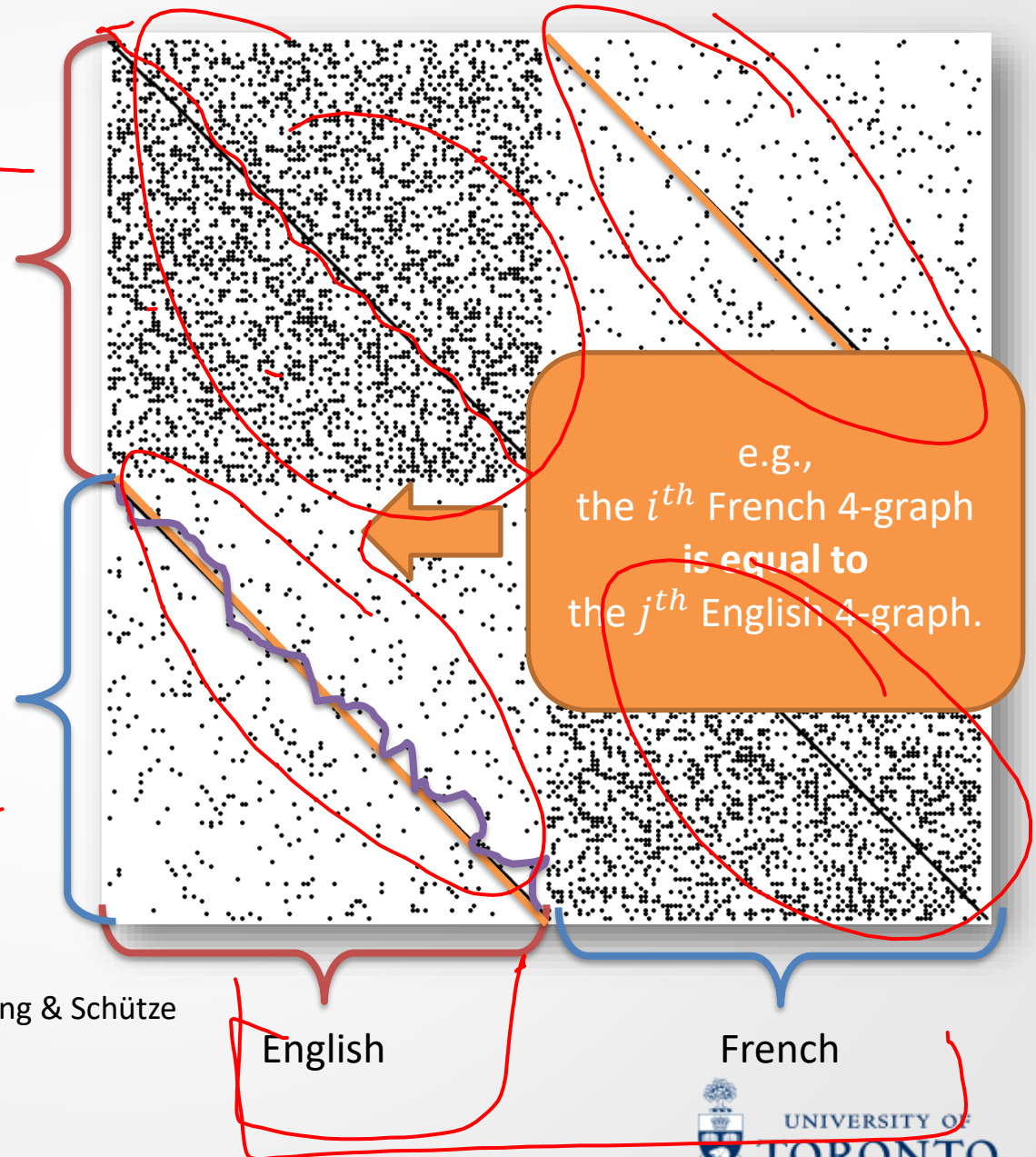
## 2. Sentence alignment by cognates

- 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**.

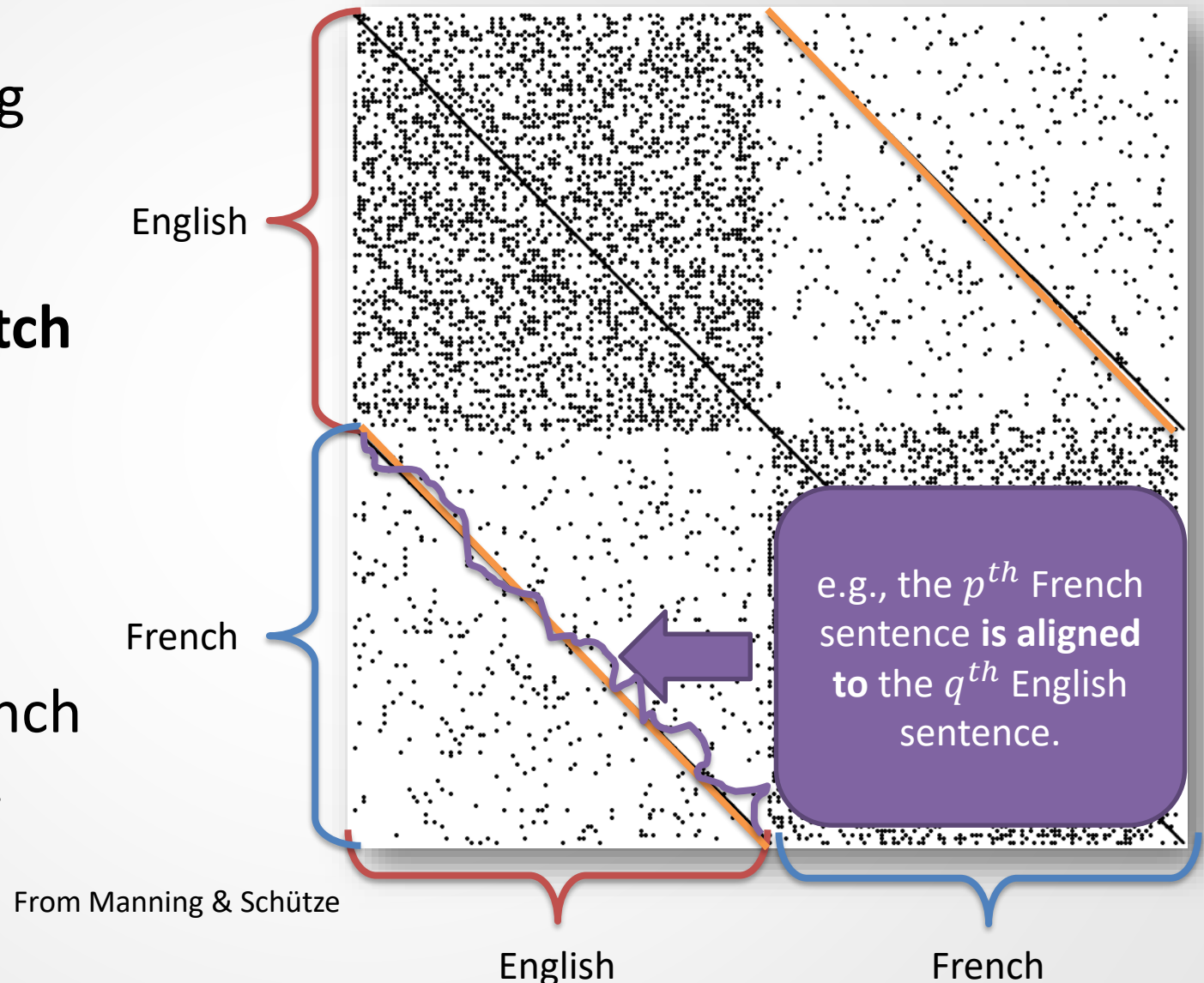
From Manning & Schütze





## 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**.



## 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.,

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

'LCS Ratio'

## 2b. Melamed's method

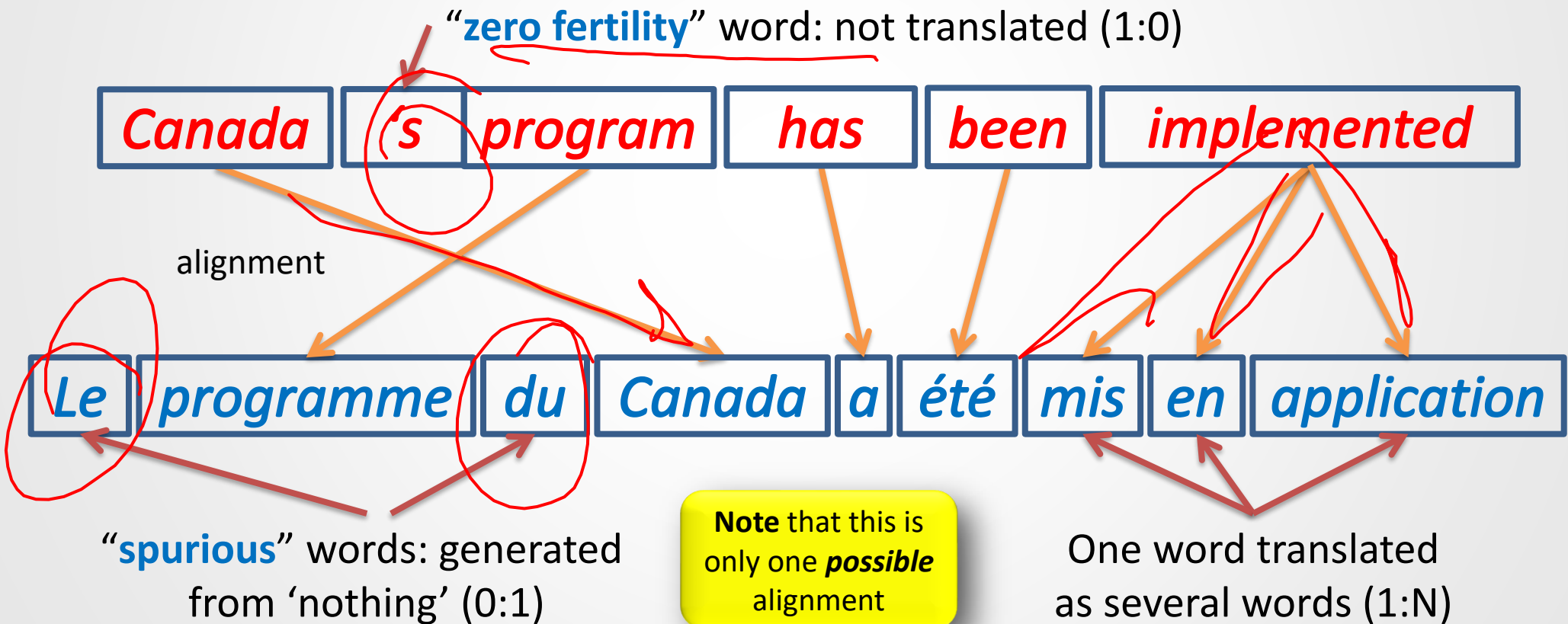
- Excludes **stop words** from both languages.  
(e.g., *the*, *a*, *le*, *un*)
- Melamed empirically declared that cognates occur when  $LCSR \geq 0.58$  (i.e., there's a lot of overlap in those words).
  - $\therefore$  25% of words in Canadian Hansard are cognates.
- As with Church, construct a “bitext” **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

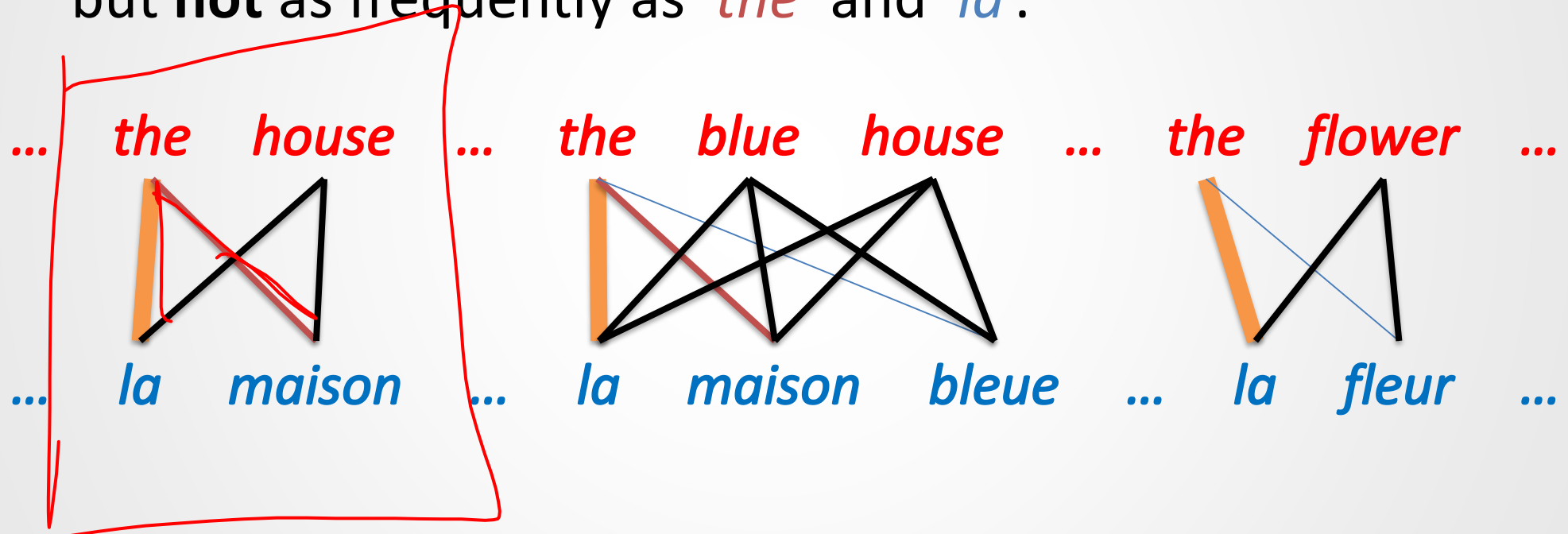
- **Word alignments** can be 1:1, N:1, 1:N, 0:1, 1:0,... E.g.,





# 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....

# 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 2<sup>nd</sup> edition of the Jurafsky & Martin text.
  - 1<sup>st</sup> 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