# Agenda for today

- Introduction to bi-text parsing
  - Review of finite-state transducers
  - Inversion transduction grammars (ITG)
  - Synchronous context-free grammars (SCFG)
- Machine translation
- Text-Based Language Processing Systems: Winter 2009

# Review finite-state string transducers

- Simultaneously generate pairs of (related) strings
- Spelling:

| | |
|---|---|
| fox | f o x |
| foxes | f o x e s |
| cat | c a t |
| cats | c a t s |
| dog | d o g |
| dogs | d o g s |
| donkey | d o n k e y |
| donkeys | d o n k e y s |

# Review finite-state string transducers

- Simultaneously generate pairs of (related) strings
- Pronunciation:

| | |
|---|---|
| fox | F AA1 K S |
| foxes | F AA1 K S AH0 Z |
| cat | K AE1 T |
| cats | K AE1 T S |
| dog | D AO1 G |
| dogs | D AA1 G Z |
| donkey | D AA1 NG K IY0 |
| donkeys | D AA1 NG K IY0 Z |

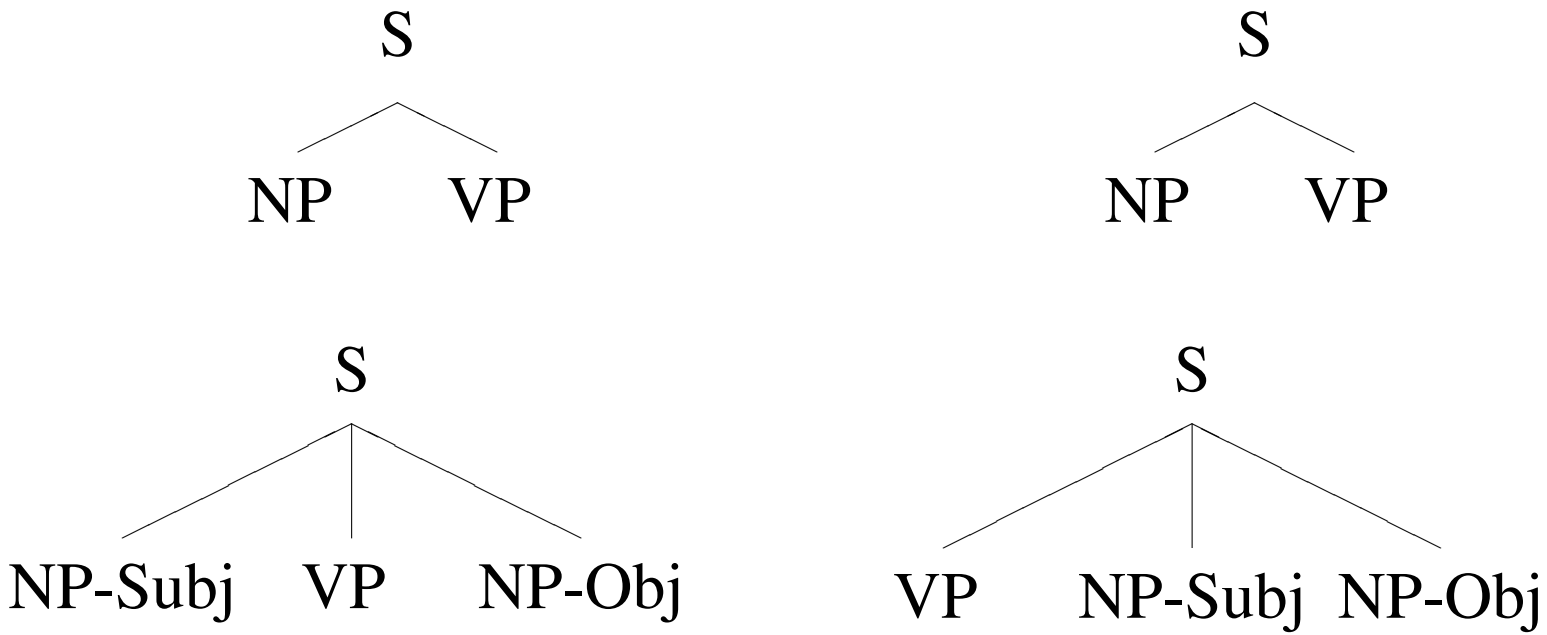
# Review finite-state string transducers

- Simultaneously generate pairs of (related) strings
- ...Translation?:

| | |
|---|---|
| fox | zorro |
| foxes | zorros |
| cat | gato |
| cats | gatos |
| mouse | ratón |
| train | formarse (v) |
| | tren (n) |
| | ir en tren (v) |
| to go by train | ir en tren (v) |

# Tree transducers

- Simultaneously generate pairs of (related) trees



## Synchronous grammars

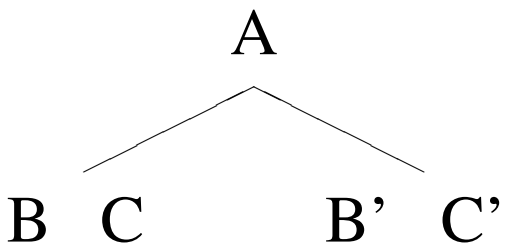
- Simultaneously generate pairs of recursively related strings
- Will always see *pairs* in the grammar rules

$$A \longrightarrow B\ C\ /\ B'\ C'$$

# Synchronous grammars

- Simultaneously generate pairs of recursively related strings
- Will always see *pairs* in the grammar rules

$$A \longrightarrow B\ C\ /\ B'\ C'$$



## Synchronous grammars as transducers

A $\longrightarrow$ fox / f o x

A $\longrightarrow$ fox / F AA1 K S

A $\longrightarrow$ fox / zorro

A $\longrightarrow$ train / formarse

A $\longrightarrow$ train / tren

## Synchronous grammars as transducers

- Drawbacks?
- Ways to address ambiguity?
- Methods similar to automata algorithms for
  - Minimization
  - Determinization
  - Weight-pushing

# Bracketing inversion transduction grammars (ITG)

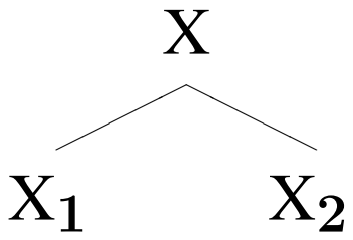
[Wu 1995,1997]

- Always binary rules
  - $X \rightarrow [\ X_1\ X_2\ ]$
  - $X \rightarrow <\ X_1\ X_2\ >$
  - $X \rightarrow s\ /\ t$
  - $X \rightarrow s\ /\ \epsilon$
  - $X \rightarrow \epsilon\ /\ t$
- Generates both source and target trees
- Requires a common binary tree
  - Derivable from parallel corpora

# ITG non-terminal rules

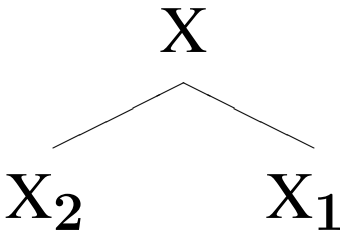
- Simple transduction production:

$$X \rightarrow [\ X_1\ X_2\ ]$$

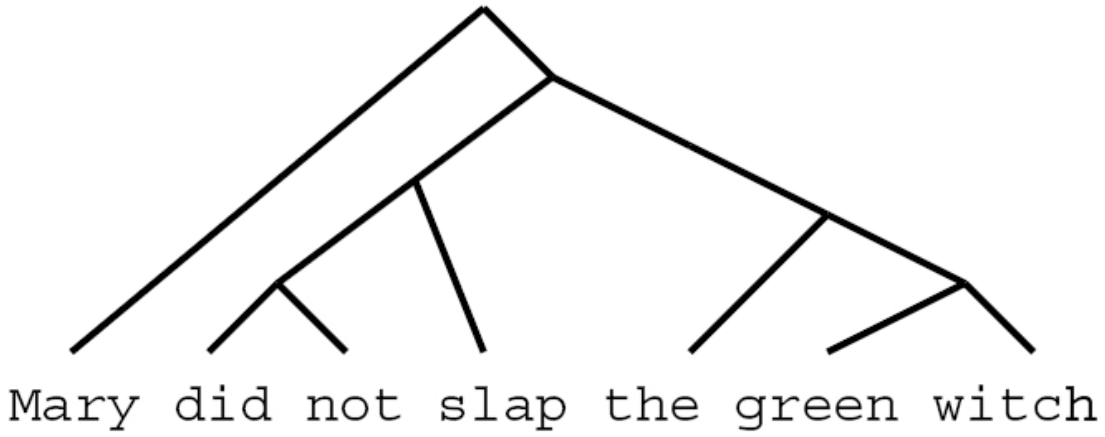



- Inverted transduction production:

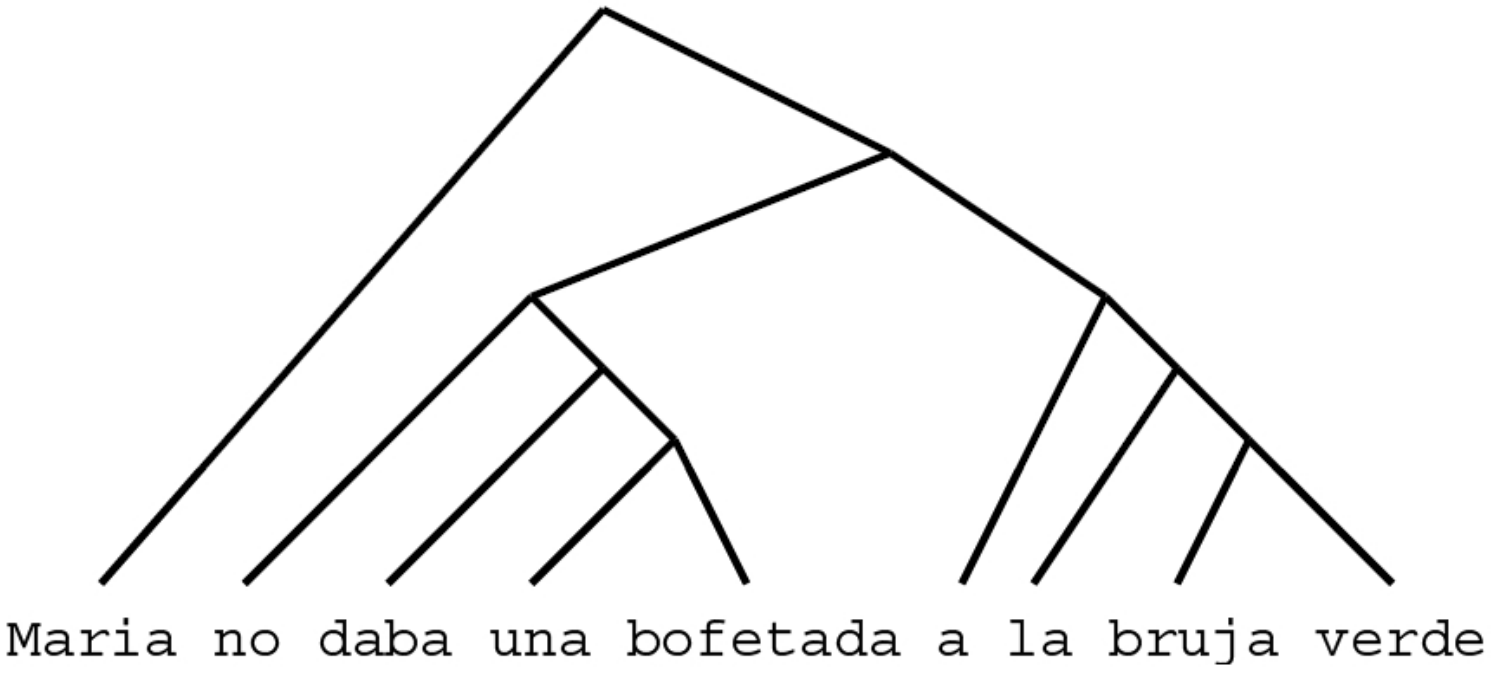
$$X \rightarrow <\ X_1\ X_2\ >$$



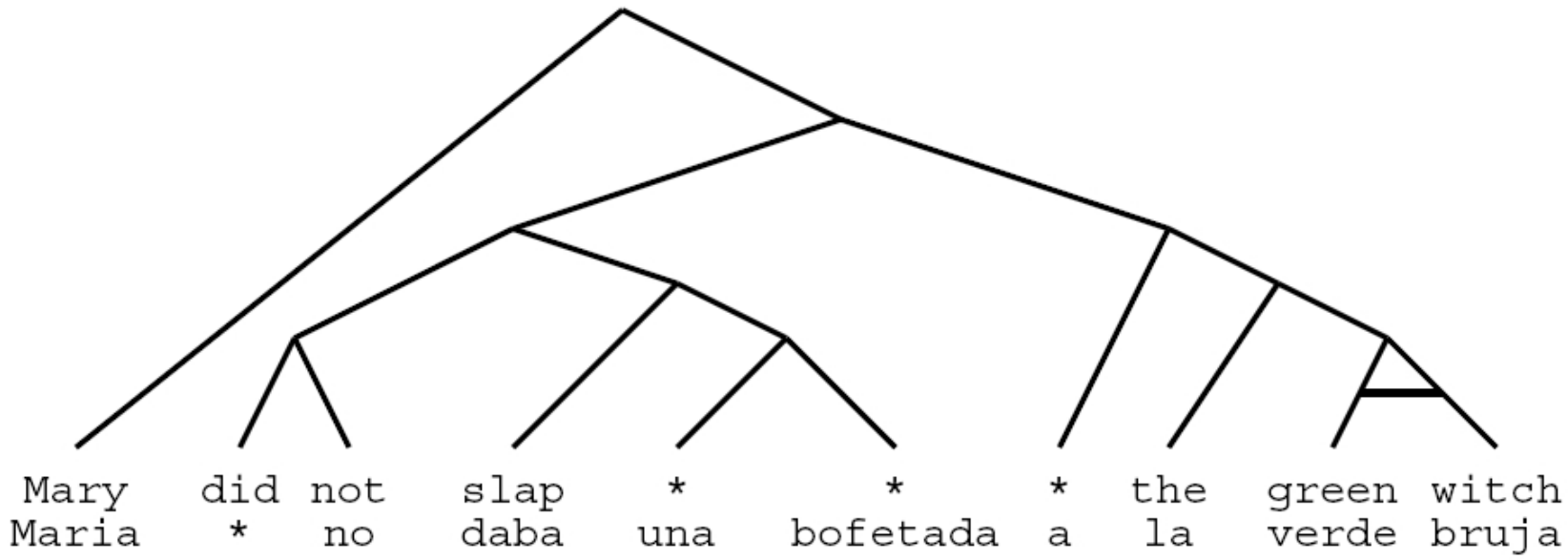
# Example: Unlabeled binary tree (English)



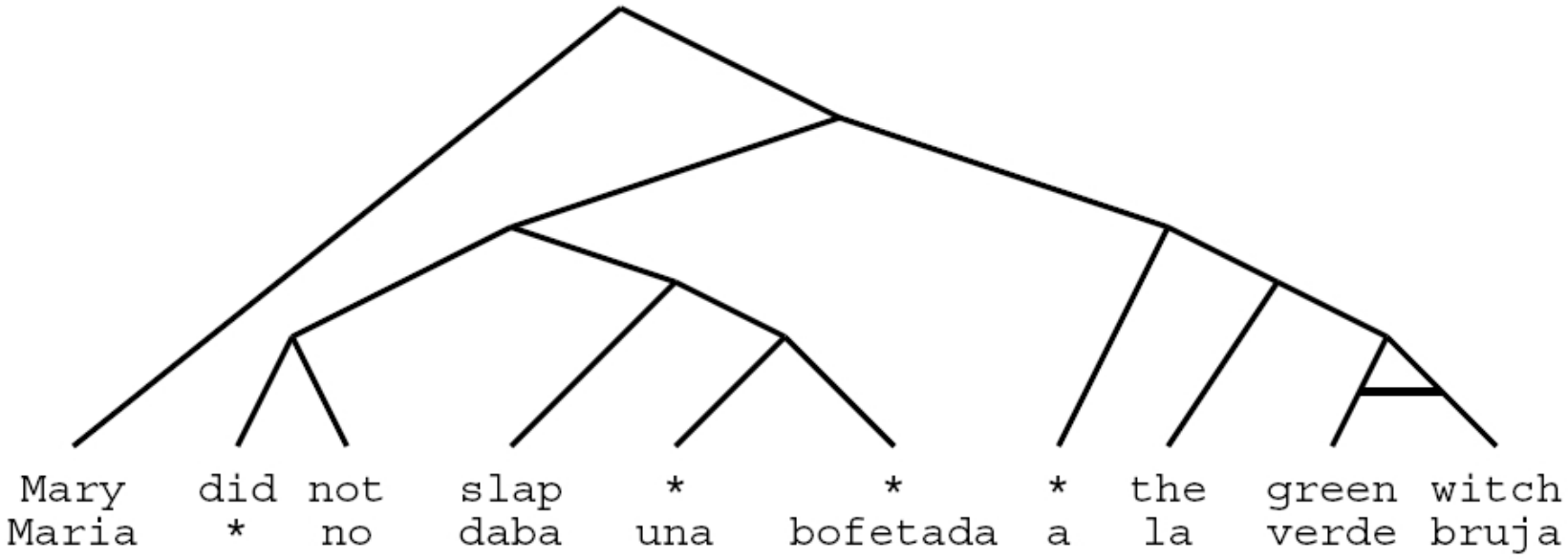
# Example: Unlabeled binary tree (Spanish)



# Example: ITG tree



# Example: ITG tree




$\mathbf{X_1 \rightarrow Mary\ /\ Maria}$

$\mathbf{X_2 \rightarrow did\ /\ \epsilon}$

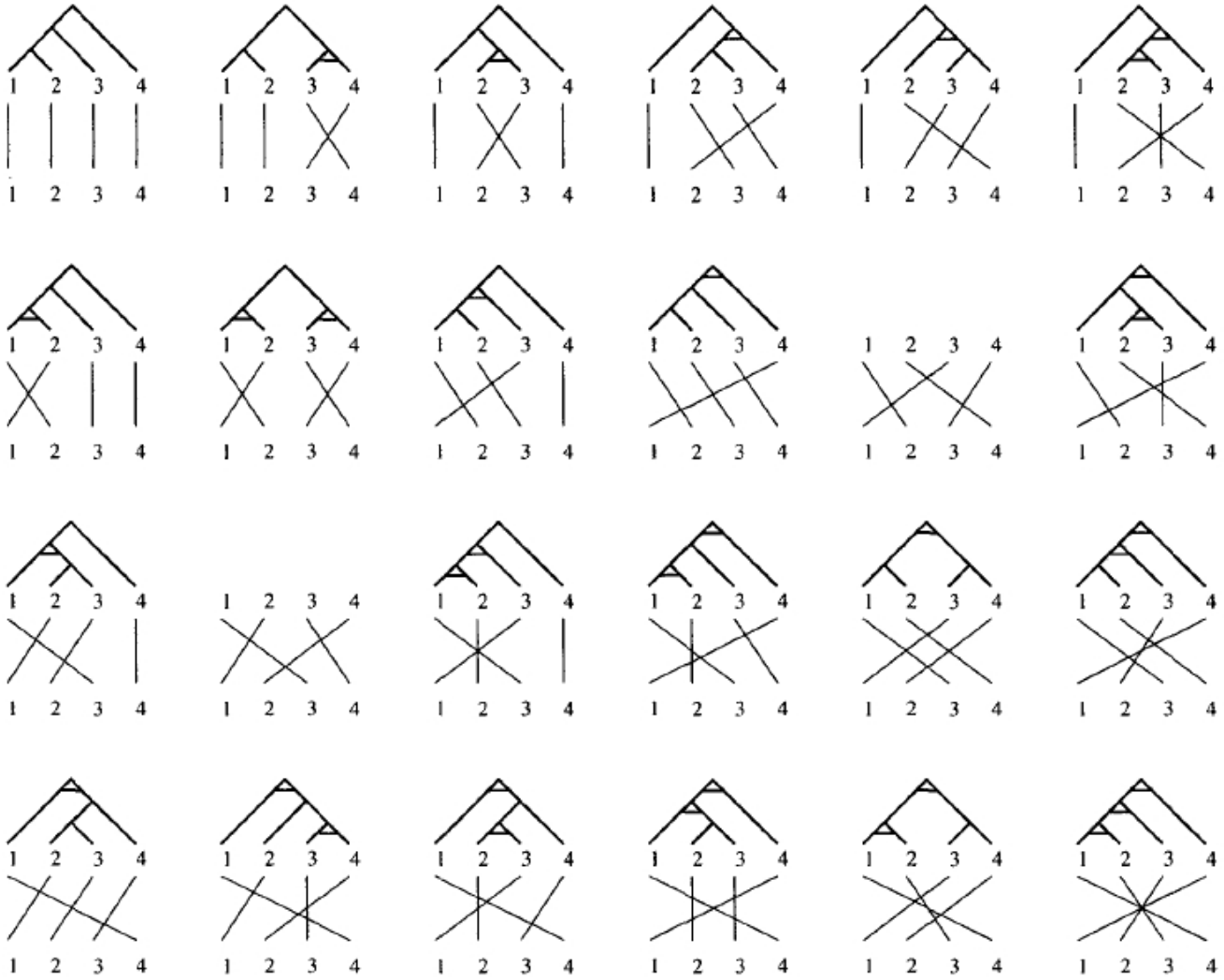
$\mathbf{...}$

$\mathbf{X_9 \rightarrow green\ /\ verde}$

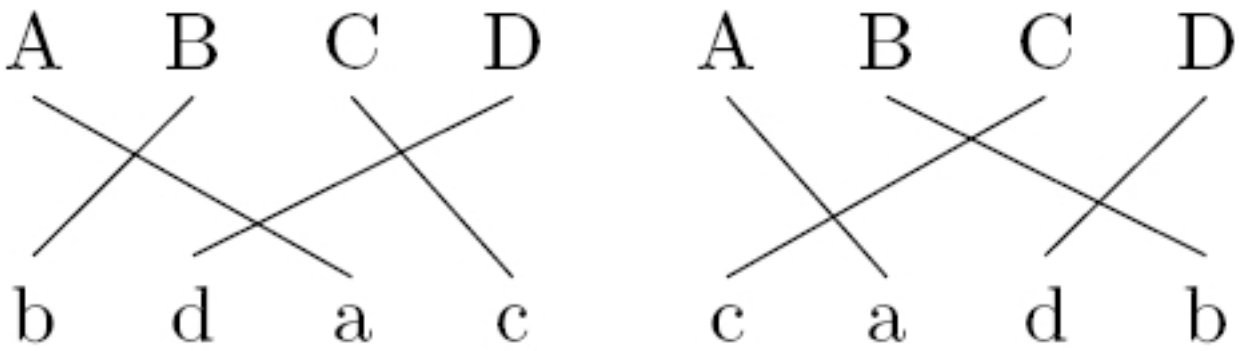
$\mathbf{X_{10} \rightarrow witch\ /\ bruja}$

$\mathbf{X_{11} \rightarrow < X_9\ X_{10} >}$

# Examples of ITG trees & alignments



# Crossing alignments




- **Impossible to represent in ITG – no corresponding trees**

# ITG summary

- **Drawbacks?**
  - **Induced structure is not (necessarily) linguistically motivated**
  - **Induced from parallel corpora**

    **(no existing human-annotated bracketed treebanks)**
  - **Similarity of two different languages' grammatical structure**
- **Benefits?**
  - **Adds *some* structure to translation**

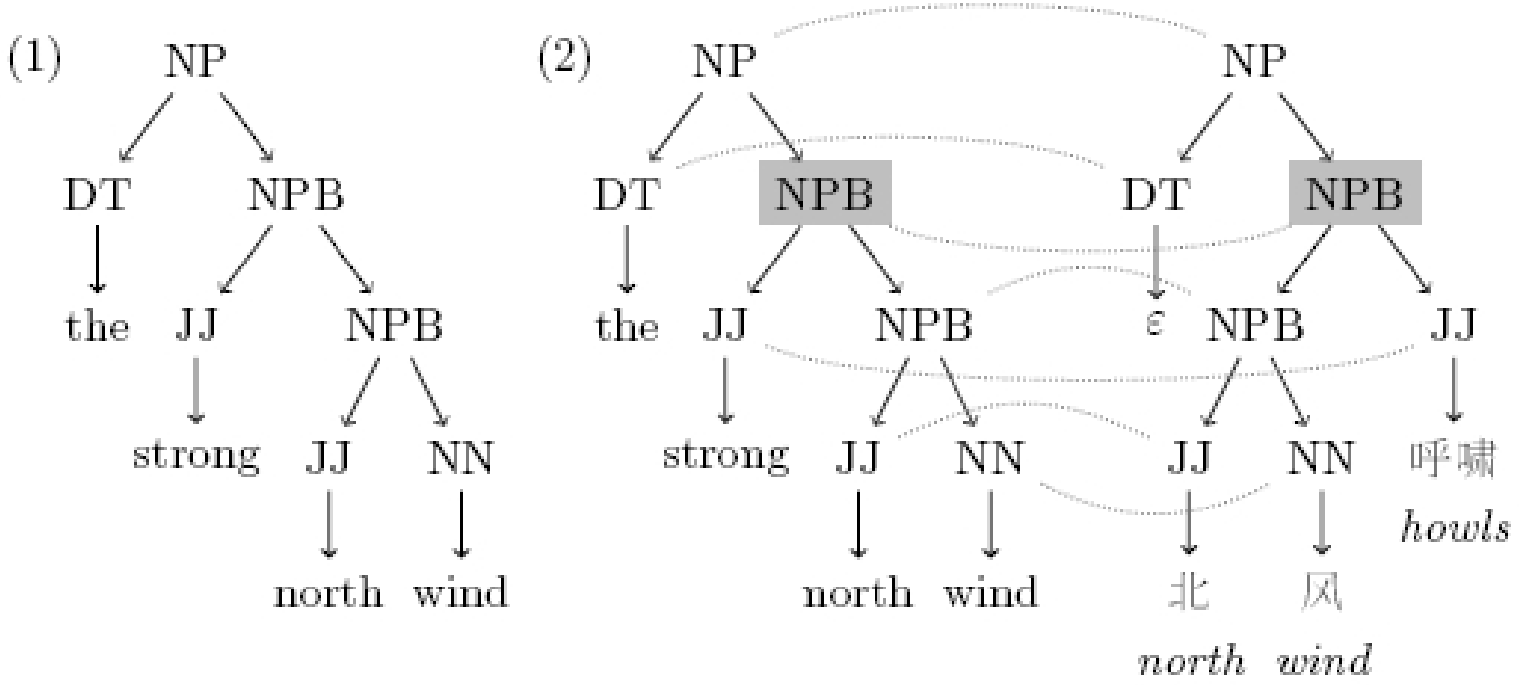
## Synchronous context-free grammars (SCFG)

- Similar to ITG...
  - Binary rules
  - Inversion productions
- Size of the non-terminal set $> 1$

  a.k.a., labeled nodes on the parse trees

$$\text{NP} \longrightarrow [\ \text{DT NPB}\ ] \qquad \text{NPB} \longrightarrow <\ \text{NPB JJ}\ >$$
$$\text{NPB} \longrightarrow [\ \text{JJ NN}\ ] \qquad \text{DT} \longrightarrow \text{the} \,/\, \epsilon$$

# Example SCFG pair of trees



## SCFG summary

- **Drawbacks?**
  - **No existing (human-annotated) corpora**
  - **Relies strongly on similarity of two languages' grammatical structure**
- **Benefits?**
  - **Linguistically motivated productions**

# Hierarchical synchronous context-free grammars

**[Chiang 05]**

- **Context-free bi-grammar**
- **Single non-terminal symbol (X)**
- **RHS of rules may include *both* non-terminal and terminals (words)**
  - $\mathrm{X} \rightarrow \mathrm{X}_1\ s_1\ \mathrm{X}_2\ /\ t_1\ \mathrm{X}_1\ \mathrm{X}_2$
- **Makes translating equivalent to parsing (though $O(n^6)$)**

# Types of hierarchical translation rules

Non-terminal rules:

S $\rightarrow$ S X / S X

X $\rightarrow$ X / X

Terminal rules:

house / casa

blue / bleu

X $\rightarrow$ slap / daba una bofetada

Mixed non-terminal/terminal:

X $\rightarrow$ not X / ne X pas

X $\rightarrow$ $X_1$ 's $X_2$ / $X_2$ de $X_1$

X $\rightarrow$ green X / X verde

# Hierarchical SCFG summary

- **Drawbacks?**
  - **Productions are not linguistically-motivated**
  - **High computational cost**
- **Benefits?**
  - **Allows for discontiguous phrases**

# Machine translation

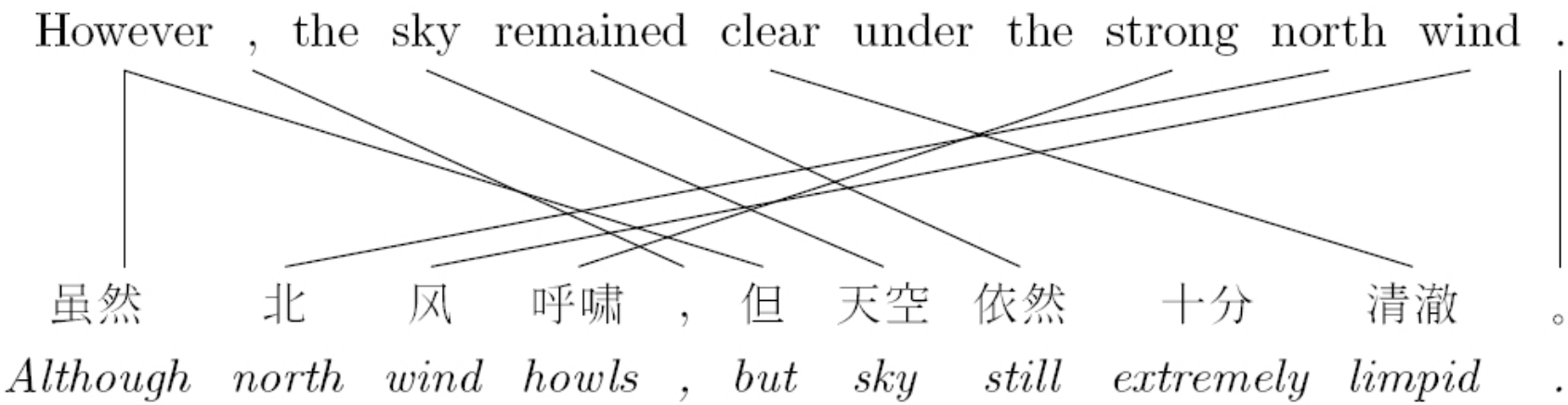
- **Word-based**
- **Phrase-based**
- **“Syntax”-based**

# Example translation

However , the sky remained clear under the strong north wind .

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 虽然 | 北 | 风 | 呼啸 | , | 但 | 天空 | 依然 | 十分 | 清澈 | 。 |
| *Although* | *north* | *wind* | *howls* | , | *but* | *sky* | *still* | *extremely* | *limpid* | . |

# Word-based alignment



# Word-based translation

- English $\longrightarrow$ Chinese:

| | |
|---|---|
| However $\longrightarrow$ *Although*$_1$ *but*$_6$ | under $\longrightarrow \epsilon$ |
| , $\longrightarrow$ ,$_5$ | the $\longrightarrow \epsilon$ |
| the $\longrightarrow \epsilon$ | strong $\longrightarrow$ *howls*$_4$ |
| sky $\longrightarrow$ *sky*$_7$ | north $\longrightarrow$ *north*$_2$ |
| remained $\longrightarrow$ *still*$_8$ | wind $\longrightarrow$ *wind*$_3$ |
| clear $\longrightarrow$ *limpid*$_{10}$ | . $\longrightarrow$ .$_{11}$ |
| | $\epsilon \longrightarrow$ *extremely*$_9$ |

- Benefits? Drawbacks?

# Phrase-based alignment

1. Divide “source” sentence into phrases (how to choose?):

   [However] [,] [the sky remained clear] [under the strong north wind] [.]

2. Translate each phrase (from a look-up table):

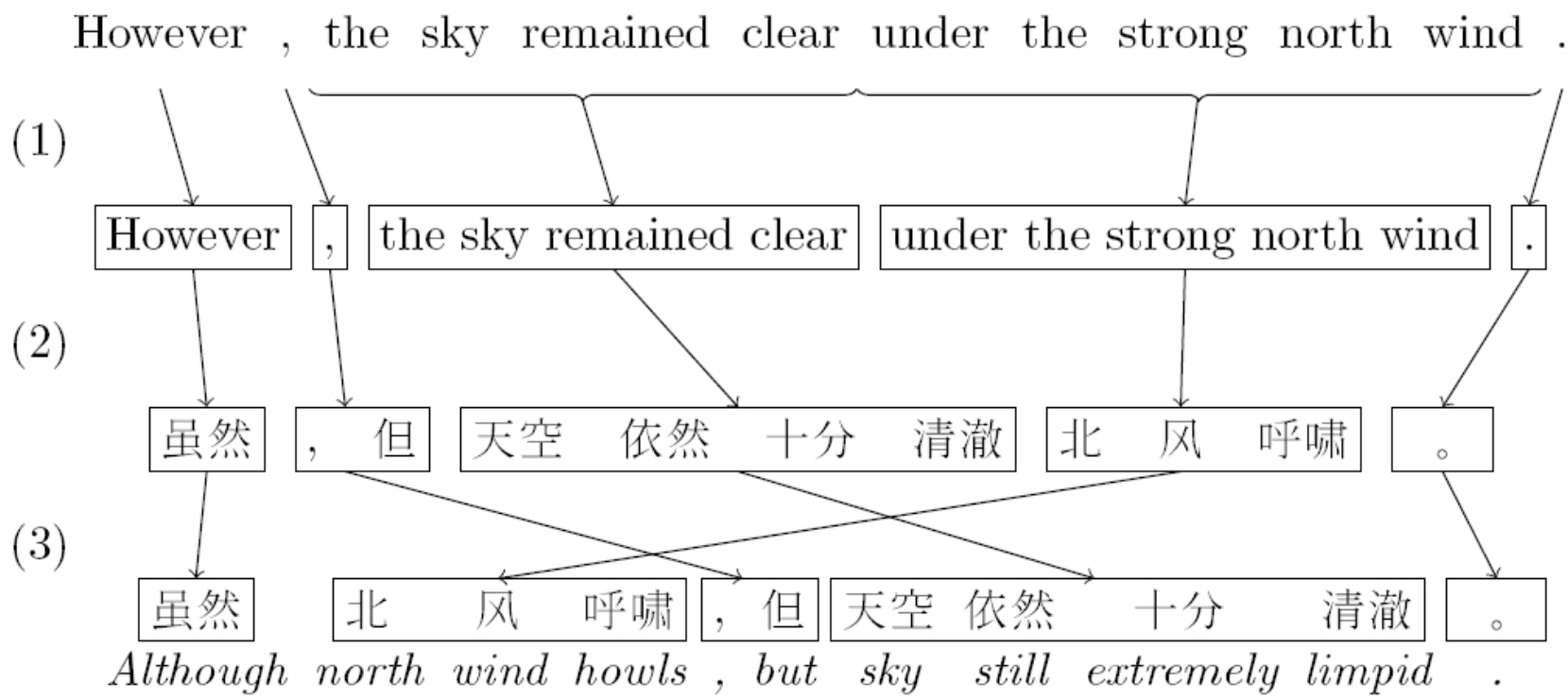
   *[Although] [, but] [sky still extremely limpid] [north wind howls] [.]*

3. Rearrange phrases (using Language Modeling?):

   *[Although] [north wind howls] [, but] [sky still extremely limpid] [.]*

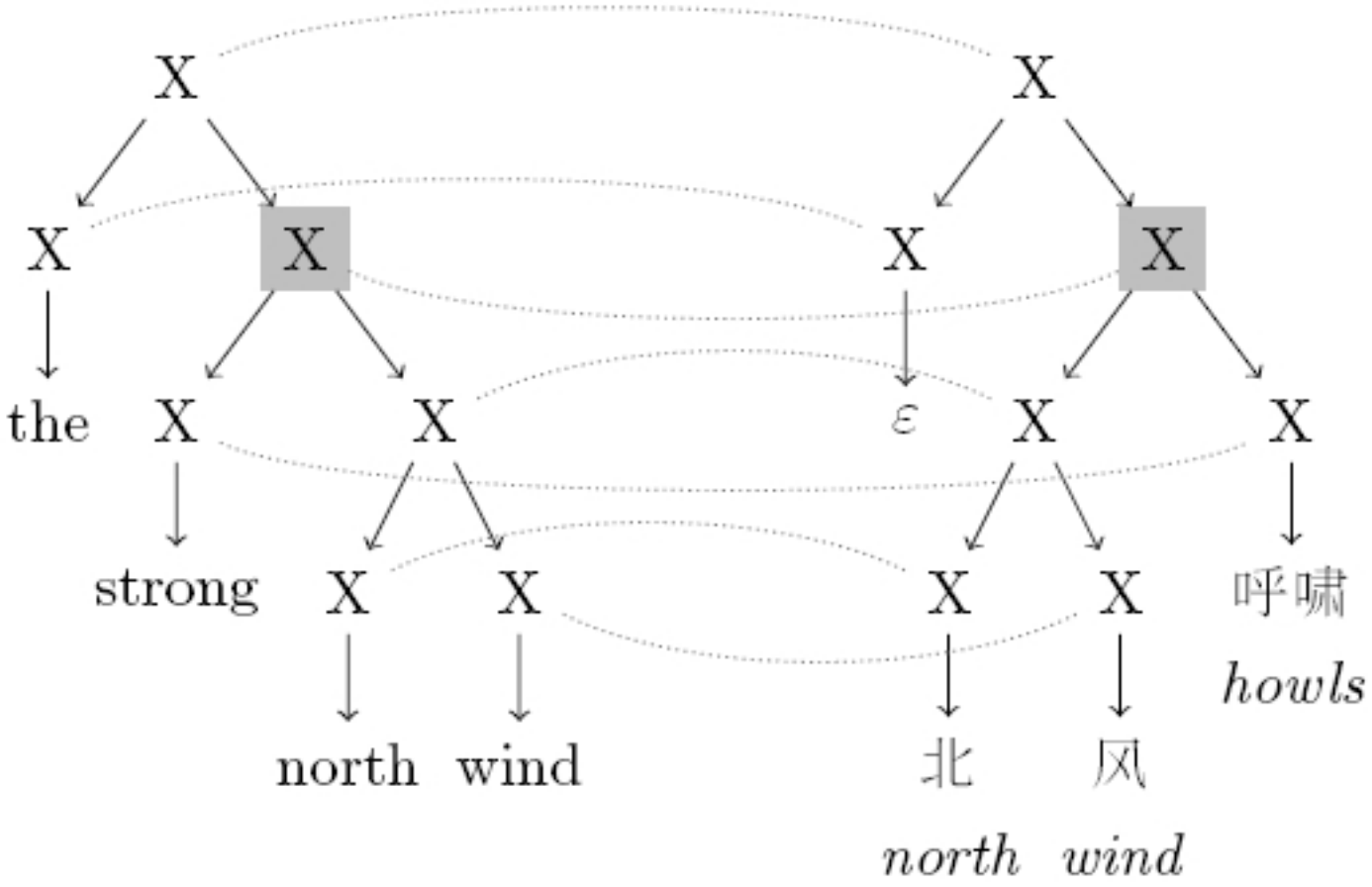
# Phrase-based translation

- English → Chinese




- Benefits? Drawbacks?

# ITG alignment



# ITG translation (derivation)

- English $\longrightarrow$ Chinese

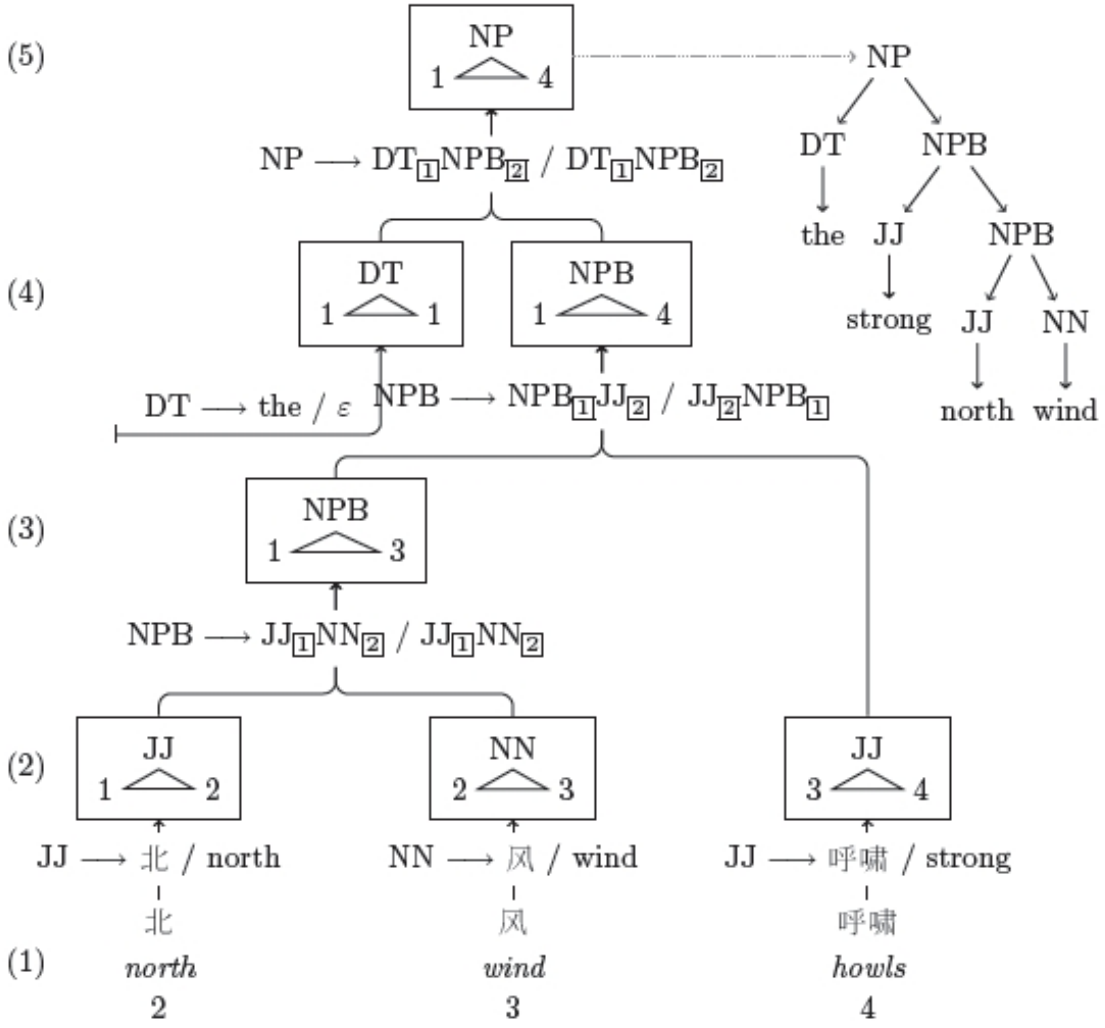
| | |
|---|---|
| X $\longrightarrow$ [ X X ] | X $\longrightarrow$ < X X > |
| X $\longrightarrow$ the / $\epsilon$ | X $\longrightarrow$ strong / *howls* |
| X $\longrightarrow$ north / *north* | X $\longrightarrow$ wind / *wind* |

- Benefits? Drawbacks?

# SCFG grammar

$$\begin{aligned}
\text{NP} &\longrightarrow \text{DT}_{\boxed{1}}\text{NPB}_{\boxed{2}} \;/\; \text{DT}_{\boxed{1}}\text{NPB}_{\boxed{2}} \\
\text{NPB} &\longrightarrow \text{JJ}_{\boxed{1}}\text{NN}_{\boxed{2}} \;/\; \text{JJ}_{\boxed{1}}\text{NN}_{\boxed{2}} \\
\text{NPB} &\longrightarrow \text{NPB}_{\boxed{1}}\text{JJ}_{\boxed{2}} \;/\; \text{JJ}_{\boxed{2}}\text{NPB}_{\boxed{1}} \\
\text{DT} &\longrightarrow \text{the} \;/\; \varepsilon \\
\text{JJ} &\longrightarrow \text{strong} \;/\; \text{呼啸} \\
\text{JJ} &\longrightarrow \text{north} \;/\; \text{北} \\
\text{NN} &\longrightarrow \text{wind} \;/\; \text{风}
\end{aligned}$$

# SCFG translation (derivation)



# Hierarchical SCFG grammar

$X \longrightarrow$ However , $X_{\boxed{1}} X_{\boxed{2}}$ . /虽然 $X_{\boxed{2}}$ , 但 $X_{\boxed{1}}$ 。

$X \longrightarrow$ under the strong north wind /北 风 呼啸

$X \longrightarrow$ the sky remained clear /天空 依然 十分 清澈

- X $\longrightarrow$ However , under the strong north wind the sky remained clear . /

  *Although north wind howls , but sky still extremely limpid .*

- X $\longrightarrow$ However , the sky remained clear under the strong north wind . /

  *Although sky still extremely limpid , but north wind howls .*

- Benefits? Drawbacks?

# Summary

- **Bi-text parsing:**

  **like finite-state transducers for trees**

- **Grammars provide more powerful transforms**
- **Higher complexity**
- **Utility of structure in translation?**

  **...syntax in translation?**