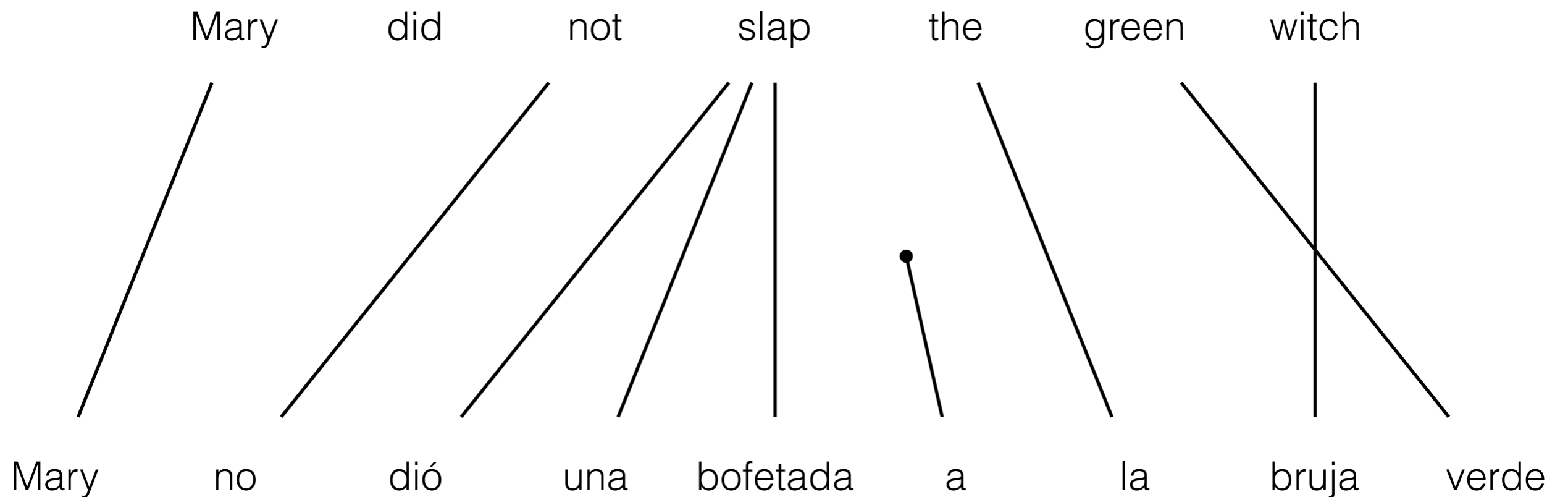


Inversion Transduction Grammars

Wilker Aziz

3/5/17

Word-based Translation



Every French word is generated by an English word (or null)

Generative Story IBM_{≥3}: Fertility

| | | | | | | | | | |
|------|----------------|-----|------|------|------|--|-----|-------|-------|
| Mary | did | not | | slap | | | the | green | witch |
| Mary | did | not | slap | slap | slap | | the | green | witch |

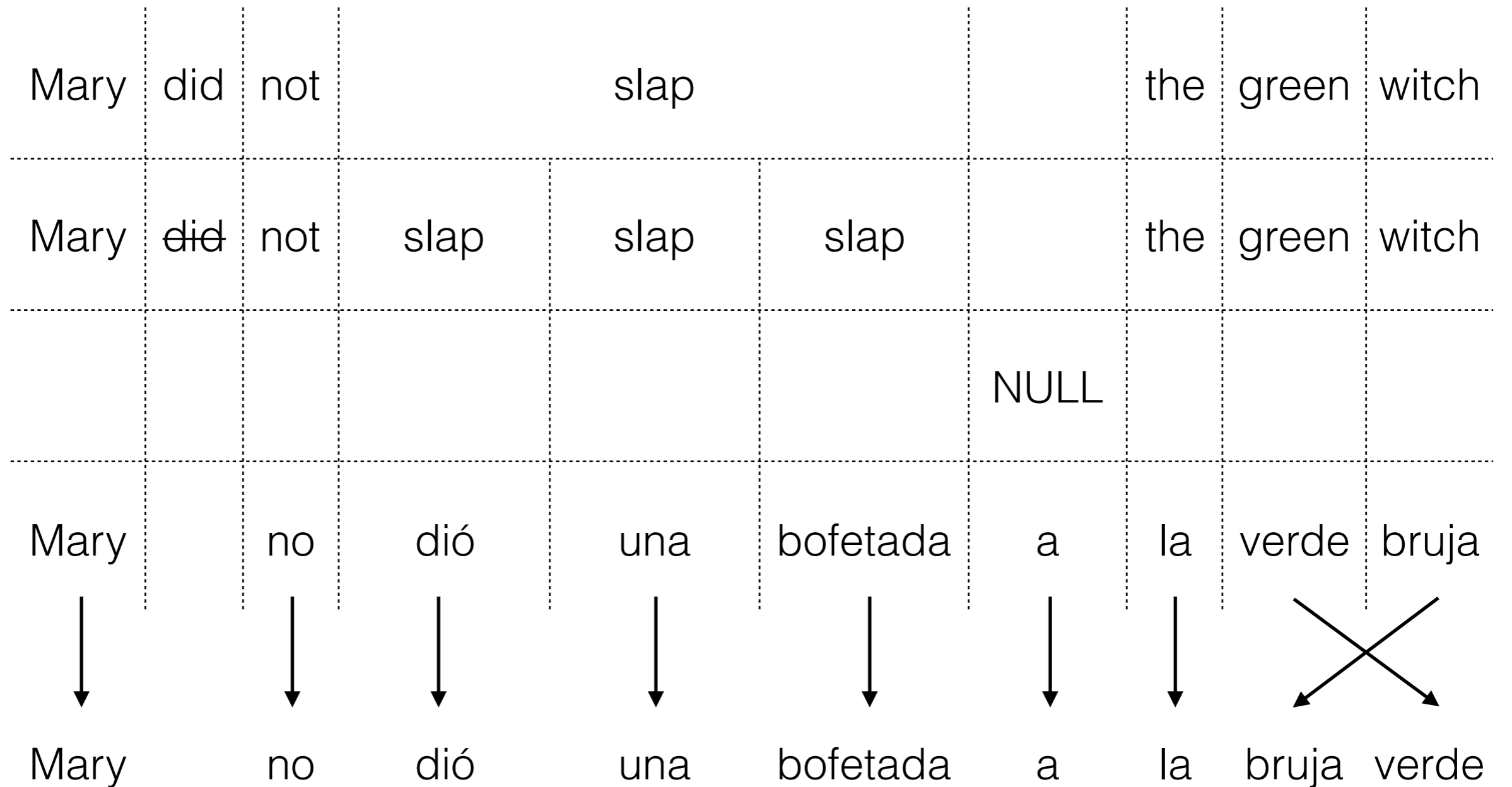
Generative Story IBM_{≥3}: NULL insertion

| | | | | | | | | | |
|------|----------------|-----|------|------|------|------|-----|-------|-------|
| Mary | did | not | | slap | | | the | green | witch |
| Mary | did | not | slap | slap | slap | | the | green | witch |
| | | | | | | NULL | | | |

Generative Story IBM_{≥3}: Translation

| | | | | | | | | | |
|------|----------------|-----|------|------|----------|------|-----|-------|-------|
| Mary | did | not | | slap | | | the | green | witch |
| Mary | did | not | slap | slap | slap | | the | green | witch |
| | | | | | | NULL | | | |
| Mary | | no | dió | una | bofetada | a | la | verde | bruja |

Generative Story IBM_{≥3}: Distortion



Discussion

- IBM models do not constrain divergence with respect to word order
- Distortion step must consider

all the $m!$ permutations

of m French words

All permutations: sensible or not?

If we do not impose structural constraints
(yet they do exist)

- the model will have to learn (rather *implicitly*)
how not to violate them
- which ought to require more data

Practical consequences

Practical consequences

Estimation

- modelling outcomes that even though possible are not plausible (unlikely to be observed)

Practical consequences

Estimation

- modelling outcomes that even though possible are not plausible (unlikely to be observed)

Generation

- NP-completeness!

NP-completeness

NP-completeness

NP-complete problem

NP-completeness

NP-complete problem

- Generalised TSP

[Knight, 1999; Zaslavskiy et al, 2009]

NP-completeness

NP-complete problem

- Generalised TSP
- Perfect matching

[Knight, 1999; Zaslavskiy et al, 2009]

[DeNero and Klein, 2008]

NP-completeness

NP-complete problem

- Generalised TSP [Knight, 1999; Zaslavskiy et al, 2009]
- Perfect matching [DeNero and Klein, 2008]
- All permutations [Asveld, 2006; 2008]

All permutations

Let $\Sigma_n = \{a_1, \dots, a_n\}$

- $S \rightarrow A_{\Sigma_n}$
- $A_X \rightarrow a A_{X-\{a\}}$ for $X \subseteq \Sigma_n, \#X \geq 2, a \in X$
- $A_{\{a\}} \rightarrow a$

Regular grammar (there is an equivalent FSA)

Complexity

Note that nonterminals are indexed by subsets of Σ_n

i.e. power set of Σ

- 2^n nonterminals (states)
- $n \times 2^n$ productions (transitions)
- $n!$ strings (paths)

Example: 3 elements

$$S \rightarrow A_{123}$$

$$A_{123} \rightarrow a_1 A_{23} \mid a_2 A_{13} \mid a_3 A_{12}$$

$$A_{12} \rightarrow a_1 A_2 \mid a_2 A_1$$

$$A_{13} \rightarrow a_1 A_3 \mid a_3 A_1$$

$$A_{23} \rightarrow a_2 A_3 \mid a_3 A_2$$

$$A_1 \rightarrow a_1$$

$$A_2 \rightarrow a_2$$

$$A_3 \rightarrow a_3$$

"IBM constraint"

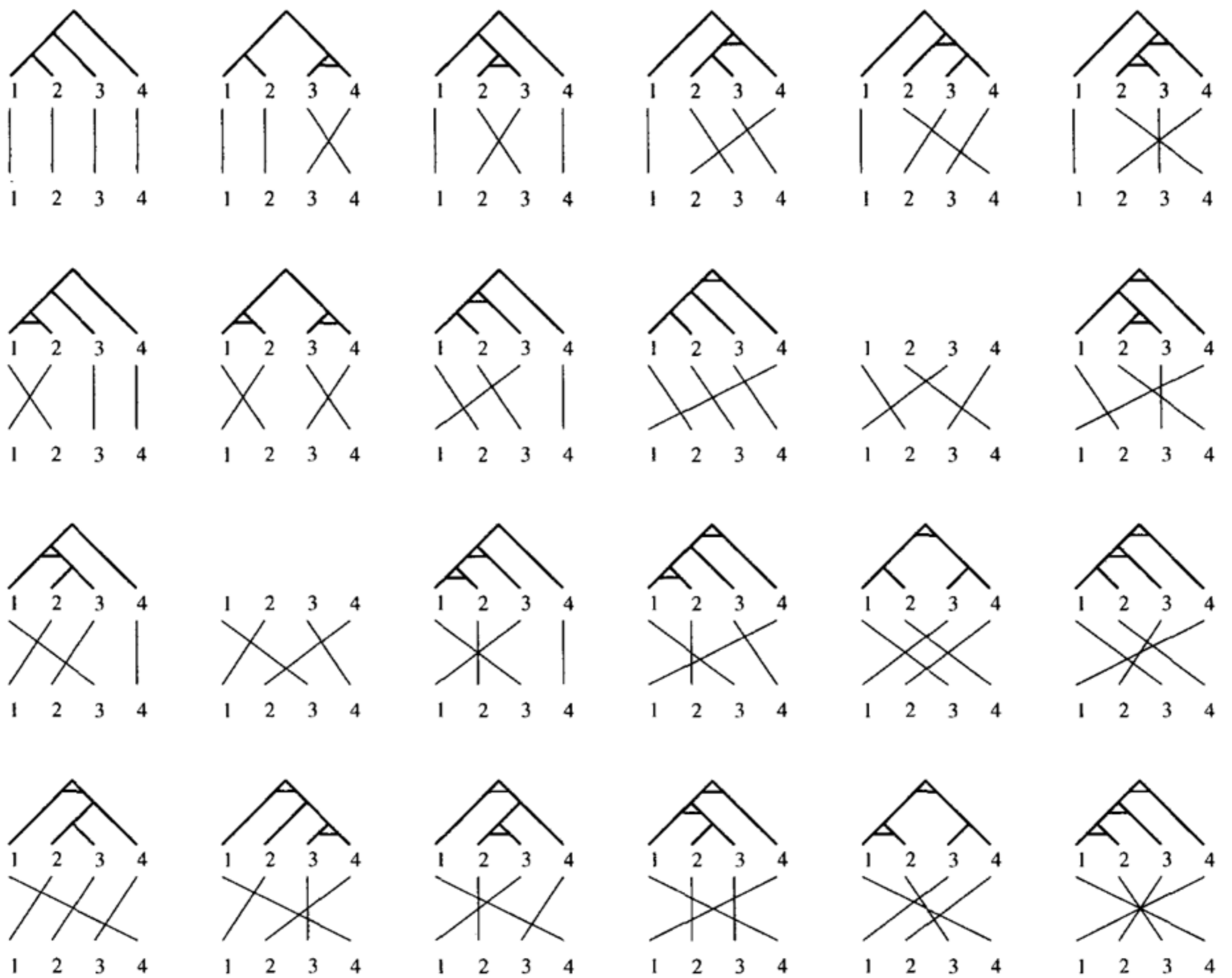
Distortion limit in **generation** but not in **estimation**

- any reasons why that may be unsatisfactory?

Constraining permutations without a distortion limit

Inversion Transduction Grammars (ITGs) [Wu, 1995; 1997]

- Binarizable permutations
 - two streams are simultaneously generated
 - context-free backbone



Number of Permutations

| r | ITG | all matchings | ratio |
|-----|---------------|--------------------|-------|
| 0 | 1 | 1 | 1.000 |
| 1 | 1 | 1 | 1.000 |
| 2 | 2 | 2 | 1.000 |
| 3 | 6 | 6 | 1.000 |
| 4 | 22 | 24 | 0.917 |
| 5 | 90 | 120 | 0.750 |
| 6 | 394 | 720 | 0.547 |
| 7 | 1,806 | 5,040 | 0.358 |
| 8 | 8,558 | 40,320 | 0.212 |
| 9 | 41,586 | 362,880 | 0.115 |
| 10 | 206,098 | 3,628,800 | 0.057 |
| 11 | 1,037,718 | 39,916,800 | 0.026 |
| 12 | 5,293,446 | 479,001,600 | 0.011 |
| 13 | 27,297,738 | 6,227,020,800 | 0.004 |
| 14 | 142,078,746 | 87,178,291,200 | 0.002 |
| 15 | 745,387,038 | 1,307,674,368,000 | 0.001 |
| 16 | 3,937,603,038 | 20,922,789,888,000 | 0.000 |

ITG

ITG

English French

ITG

| | English | French | |
|-----------------|---------|--------|------|
| $S \rightarrow$ | X | X | copy |

ITG

| | English | French | |
|-----------------|-----------|-----------|------|
| $S \rightarrow$ | X | X | copy |
| $X \rightarrow$ | $X_1 X_2$ | $X_1 X_2$ | copy |

ITG

| | English | French | |
|-----------------|-----------|-----------|--------|
| $S \rightarrow$ | X | X | copy |
| $X \rightarrow$ | $X_1 X_2$ | $X_1 X_2$ | copy |
| | | $X_2 X_1$ | invert |

ITG

| | English | French | |
|-----------------|-----------|-----------|-----------|
| $S \rightarrow$ | X | X | copy |
| $X \rightarrow$ | $X_1 X_2$ | $X_1 X_2$ | copy |
| | | $X_2 X_1$ | invert |
| $X \rightarrow$ | e | f | transduce |

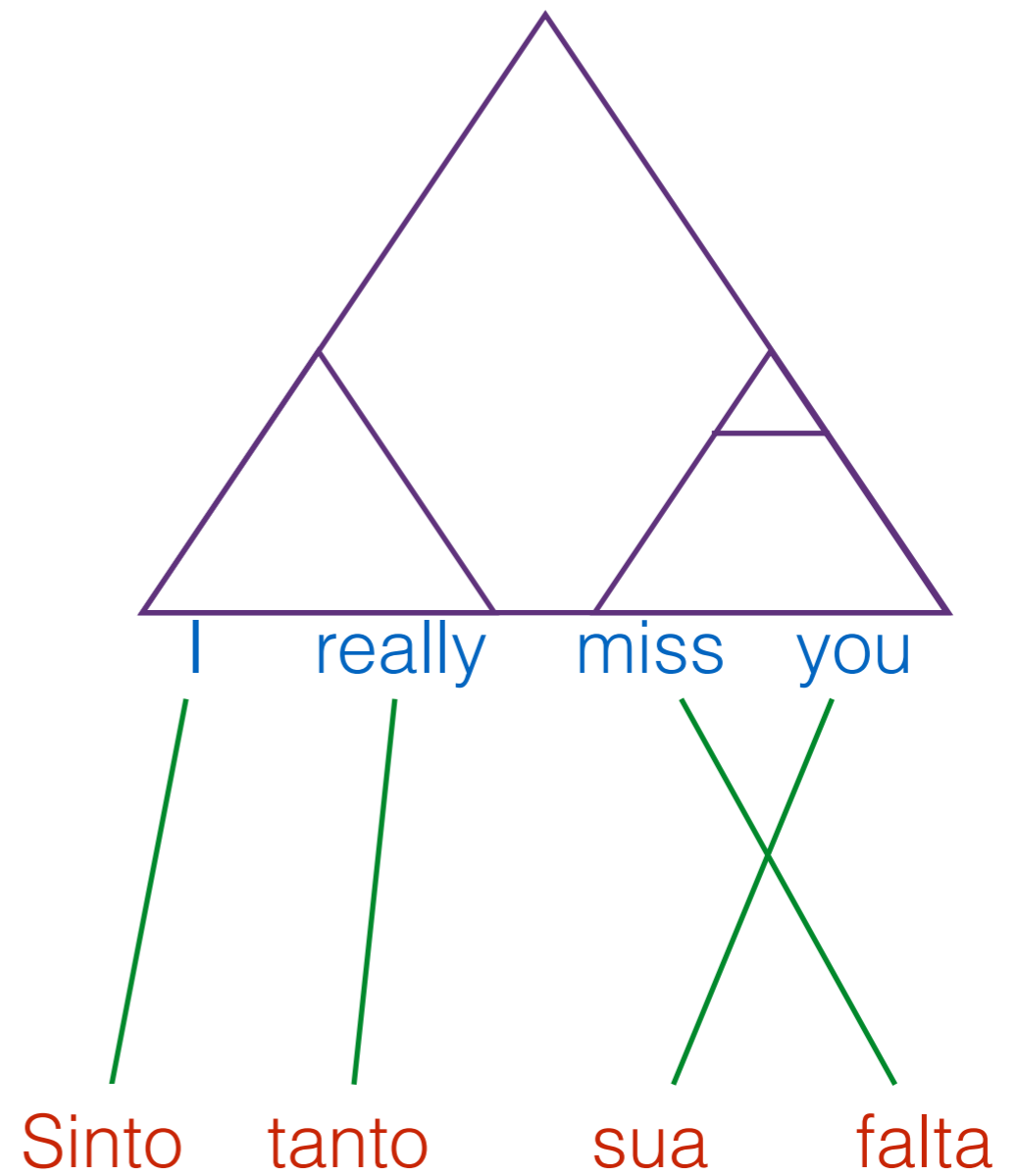
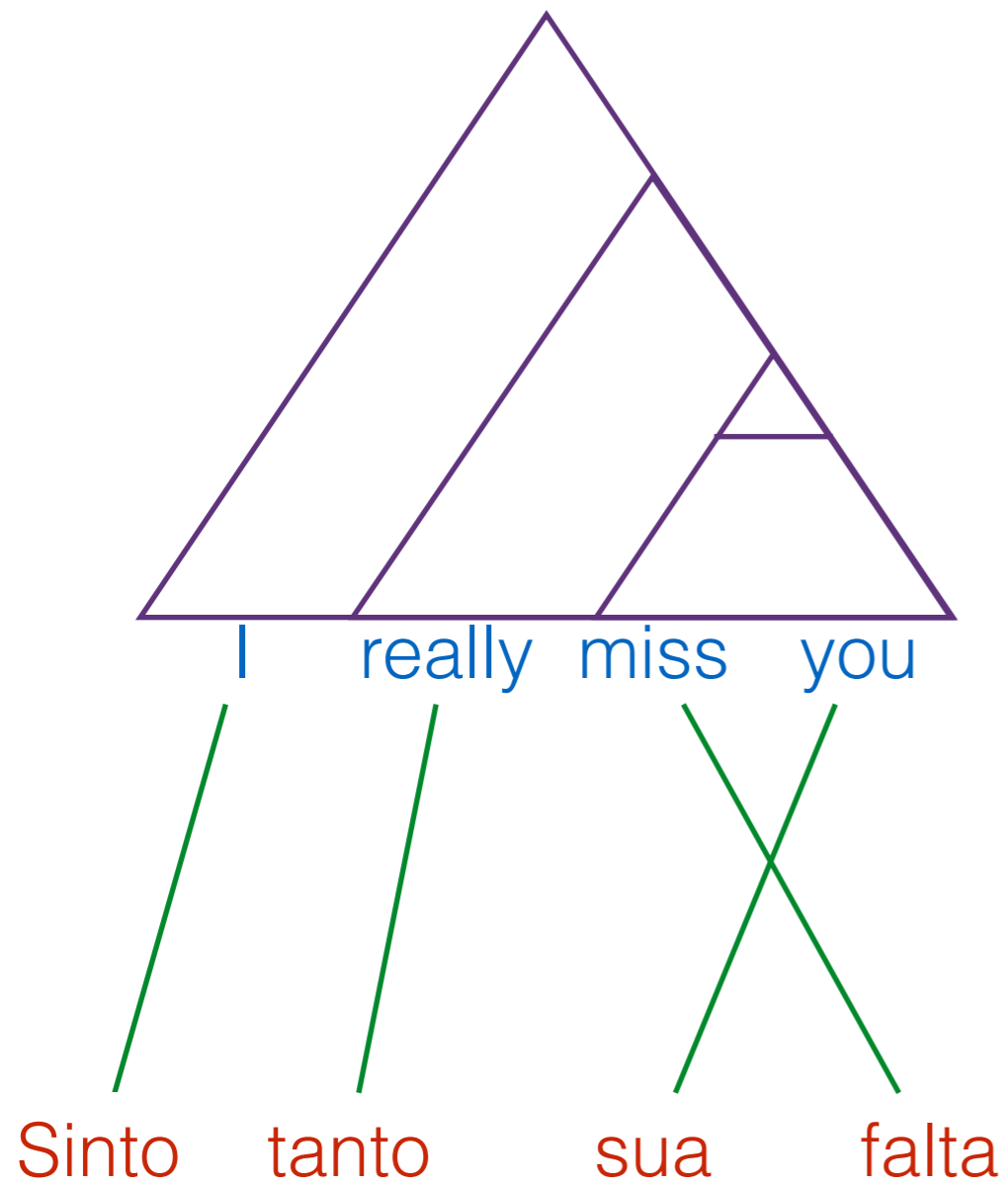
ITG

| | English | French | |
|-----------------|-----------|---------------|-----------|
| $S \rightarrow$ | X | X | copy |
| $X \rightarrow$ | $X_1 X_2$ | $X_1 X_2$ | copy |
| | | $X_2 X_1$ | invert |
| $X \rightarrow$ | e | f | transduce |
| $X \rightarrow$ | e | ε | delete |

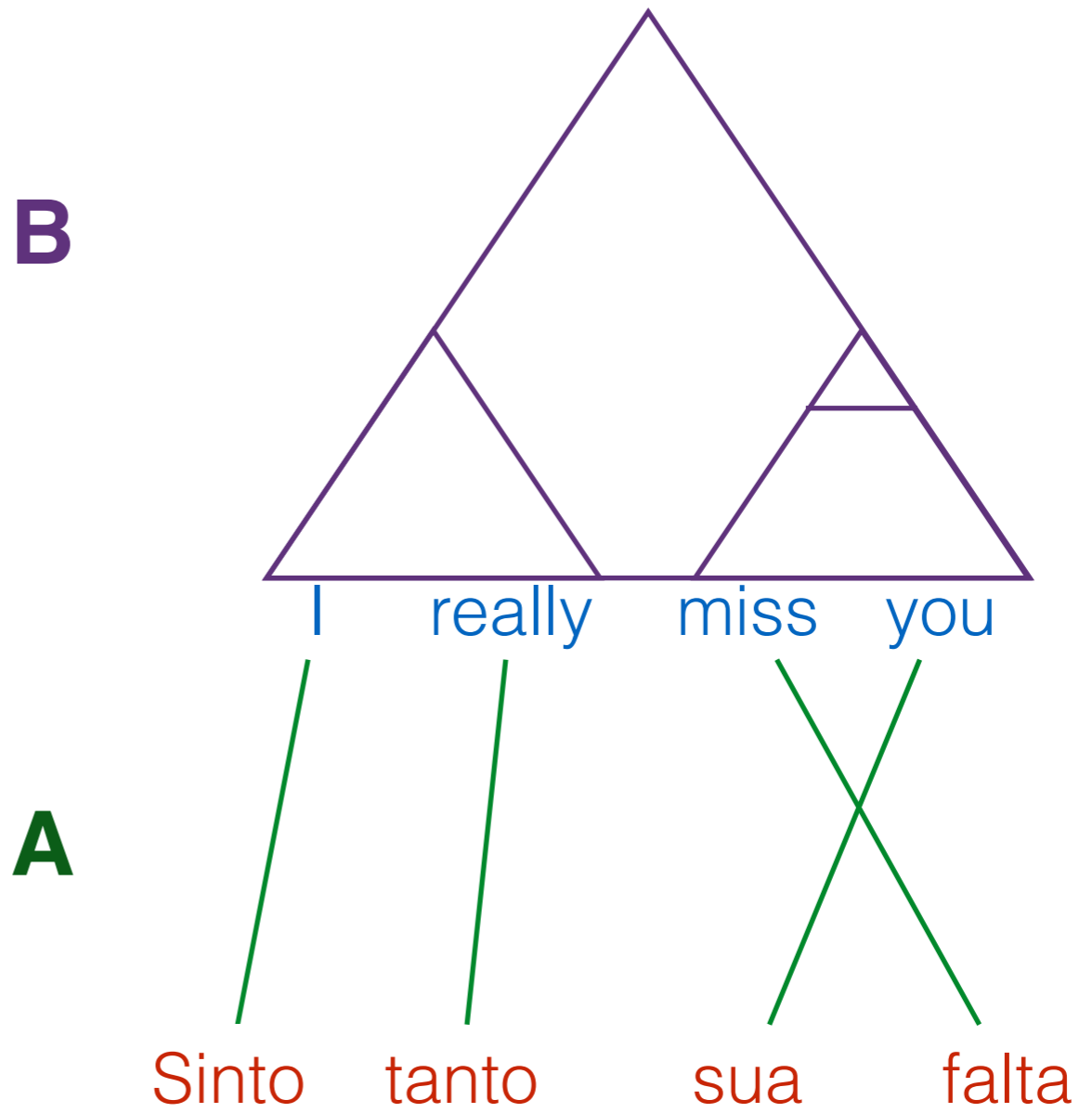
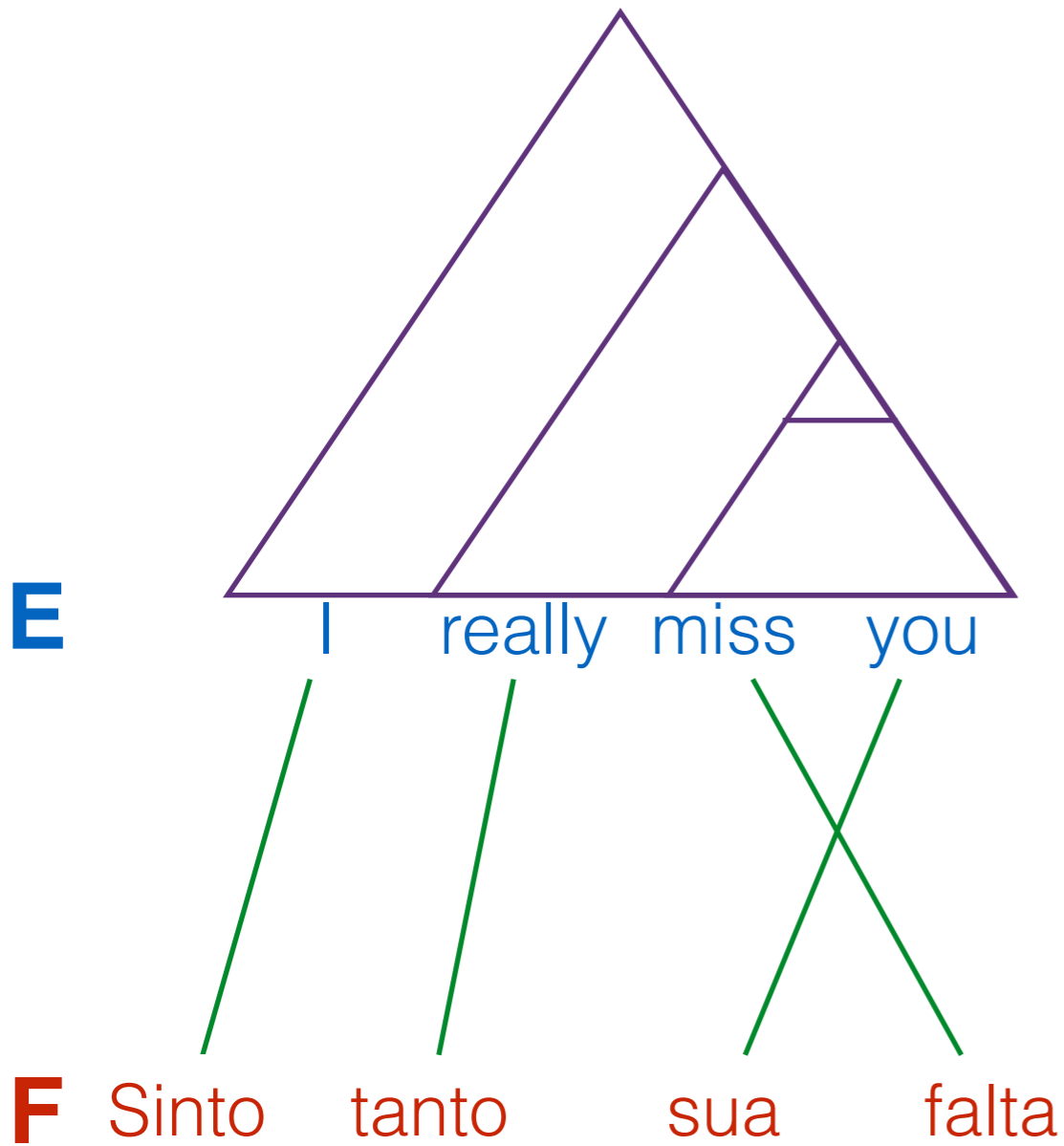
ITG

| | English | French | |
|-----------------|------------|------------|-----------|
| $S \rightarrow$ | X | X | copy |
| $X \rightarrow$ | $X_1 X_2$ | $X_1 X_2$ | copy |
| | | $X_2 X_1$ | invert |
| $X \rightarrow$ | e | f | transduce |
| $X \rightarrow$ | e | ϵ | delete |
| $X \rightarrow$ | ϵ | f | insert |

ITG Trees



ITG Trees



Model

Joint probability model $P(T) = P(A, B, E, F)$

$$t = \langle r_1, \dots, r_n \rangle$$

$$e = \text{yield}_1(t)$$

$$f = \text{yield}_2(t)$$

$$a = \text{alignment}(t)$$

$$b = \text{bracketing}(t)$$

$$P(T = t) = P(A = a, B = b, E = e, F = f)$$

$$= \prod_{i=1}^N \theta_{r_i}$$

Parametrisation

Parametrisation

Multinomial: one parameter per rule

Parametrisation

Multinomial: one parameter per rule

- θ_{\square} one parameter for **monotone**

Parametrisation

Multinomial: one parameter per rule

- θ_{\square} one parameter for **monotone**
- $\theta_{\langle \rangle}$ one parameter for **swap**

Parametrisation

Multinomial: one parameter per rule

- θ_{\square} one parameter for **monotone**
- $\theta_{\langle \rangle}$ one parameter for **swap**
- $\theta_{e/f}$ one parameter per **word pair**

Parametrisation

Multinomial: one parameter per rule

- θ_{\square} one parameter for **monotone**
- $\theta_{\langle \rangle}$ one parameter for **swap**
- $\theta_{e/f}$ one parameter per **word pair**
- $\theta_{e/\varepsilon}$ one parameter per deleted **English** word

Parametrisation

Multinomial: one parameter per rule

- θ_{\square} one parameter for **monotone**
- $\theta_{\langle \rangle}$ one parameter for **swap**
- $\theta_{e/f}$ one parameter per **word pair**
- $\theta_{e/\varepsilon}$ one parameter per deleted **English** word
- $\theta_{\varepsilon/f}$ one parameter per inserted **French** word

MLE

We do not typically construct treebanks of ITG trees

- **potential** counts instead of *observed* counts

$$\theta_{X \rightarrow \alpha} = \frac{\langle n(X \rightarrow \alpha) \rangle P(A, B | F, E)}{\sum_{\alpha'} \langle n(X \rightarrow \alpha') \rangle P(A, B | F, E)}$$

Expectations from parse forests

- Inside-Outside [Baker, 1979; Lari and Young, 1990; Goodman, 1999]

Typically initialised with IBM1

Difficulties

Inference: complexity $O(l^3m^3)$

Model: too few reordering parameters

Decisions: ambiguity

- Disambiguation problem is NP-complete [Sima'an, 1996]

$$\arg \max_A P(A|F, E) = \arg \max_A \sum_B P(A, B|F, E)$$

$$\approx \arg \max_{A, B} P(A, B|F, E)$$

Bibliography

- Knight, Kevin. 1999. Decoding complexity in word-replacement translation models. In *Computational Linguistics*. MIT Press.
- Zaslavskiy, Mikhail and Dymetman, Marc and Cancedda, Nicola. 2009. Phrase-based statistical machine translation as a traveling salesman problem. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*.
- DeNero, John and Klein, Dan. 2008. The Complexity of Phrase Alignment Problems. In *Proceedings of ACL-08: HLT*.
- Asveld, Peter R. J. 2006. Generating All Permutations by Context-free Grammars in Chomsky Normal Form. In *Theoretical Computer Science*. Elsevier Science Publishers Ltd.
- Asveld, Peter R. J. 2008. Generating All Permutations by Context-free Grammars in Greibach Normal Form. In *Theoretical Computer Science*. Elsevier Science Publishers Ltd.
- Wu, D. 1995. An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. ACL.

Bibliography

- Wu, D. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. In *Computational Linguistics*. MIT Press.
- James K. Baker. 1979. Trainable grammars for speech recognition. In *Proceedings of the Spring Conference of the Acoustical Society of America*.
- Karim Lari and Steve J. Young. 1990. The estimation of stochastic context-free grammars using the inside--outside algorithm. In *Computer Speech and Language*.
- Goodman, Joshua. 1999. Semiring parsing. In *Computational Linguistics*.
- Sima'an, Khalil. 1996. Computational complexity of probabilistic disambiguation by means of tree-grammars. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*.