

# Introduction to Statistical Machine Translation

Kenji Yamada  
Xerox Research Centre Europe

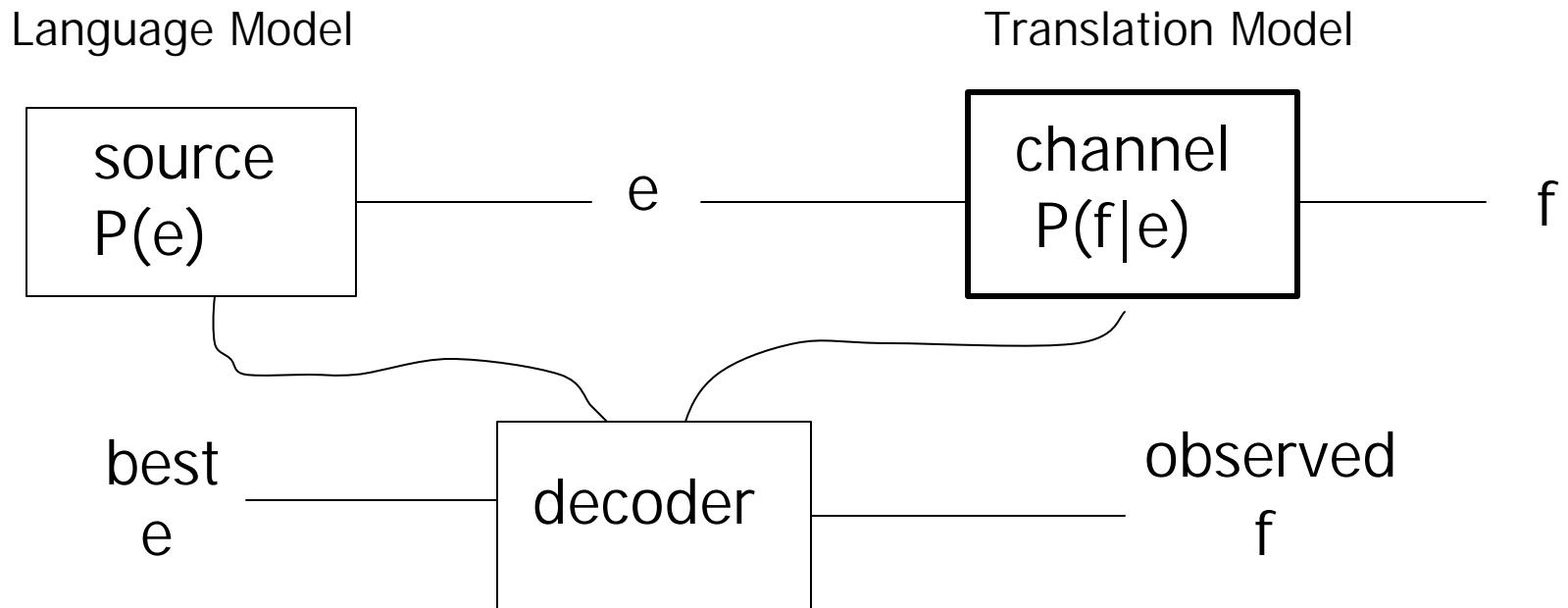
# What is Statistical MT?

- Traditional MT = rule-based  
Human written (several years)
- Statistical MT = data-driven  
Statistical Model  
Parameter estimation (learn from input/output pairs)  
Translation = decoding

# Statistical MT as ...

- Instance of Machine Learning problem
  - Learn function of French  $\rightarrow$  English
- A kind of Speech Recognition
  - Audio signal  $\rightarrow$  word sequence
  - Noisy channel model

# Noisy Channel Model



$$\operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e)P(e)$$

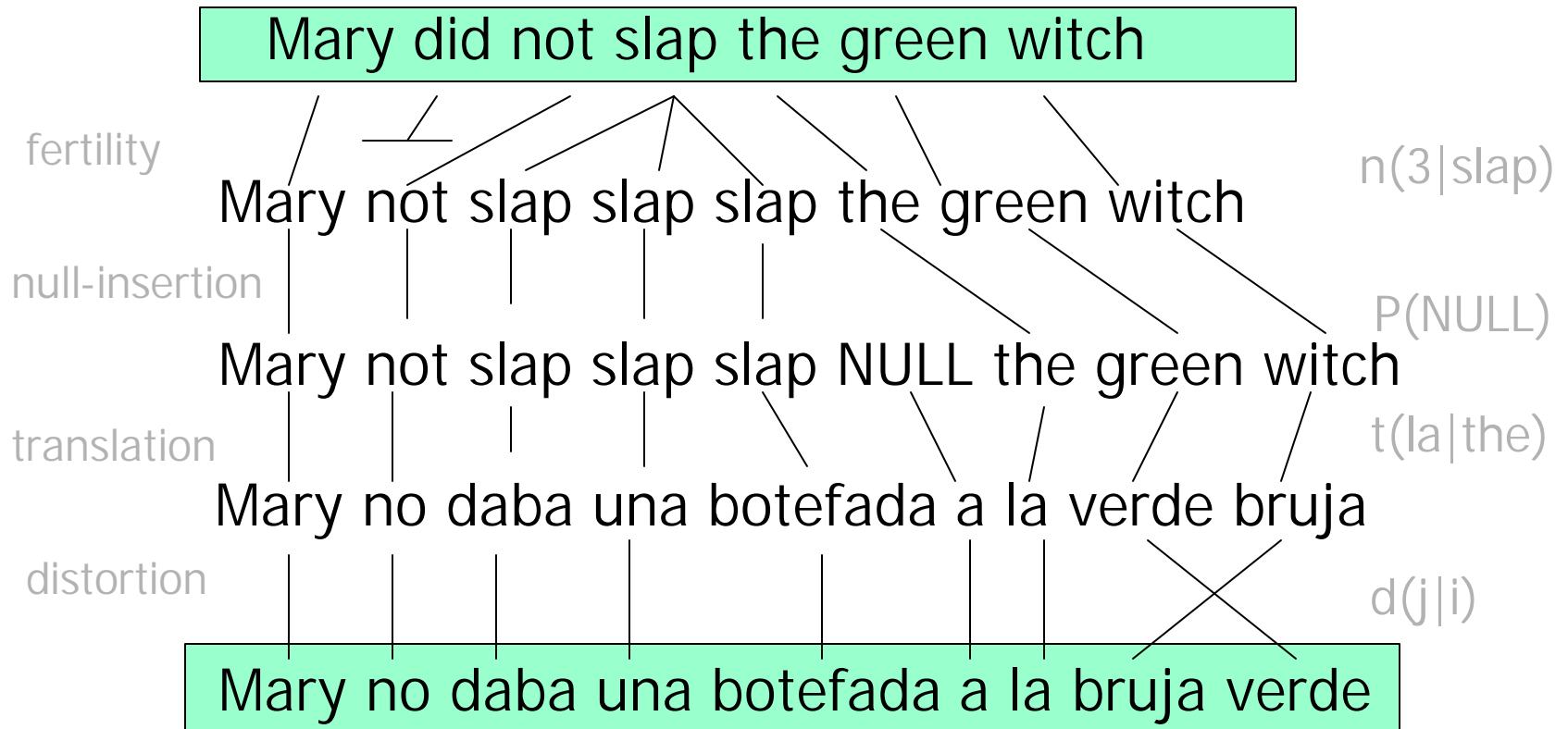
# Decompose a complex problem

- Traditional (rule-based) MT
  - Analyze and generate
  - Morphology, syntax, semantics, ...
- Statistical MT
  - Mathematically easy decomposition
  - Utilize existing parameter estimation algorithm
  - Simple model, huge training data  
(rely on computational power)

# Translation Models

- Word-based Models
  - IBM Model (model 1-5) [Brown, et al. 1993]
- Phrase-based Models
  - Wang's model [Wang and Waibel, 1998]
  - Alignment Templates [Och et al., 1999]
- Syntax-based Models
  - Inversion Transduction Grammar [Wu, 1997]
  - Head Automata [Alshawi et al., 2000]
  - Tree-to-string model [Yamada and Knight, 2001]
  - Tree-to-tree models [Hajic et al, 2002], [Glidea 2003]

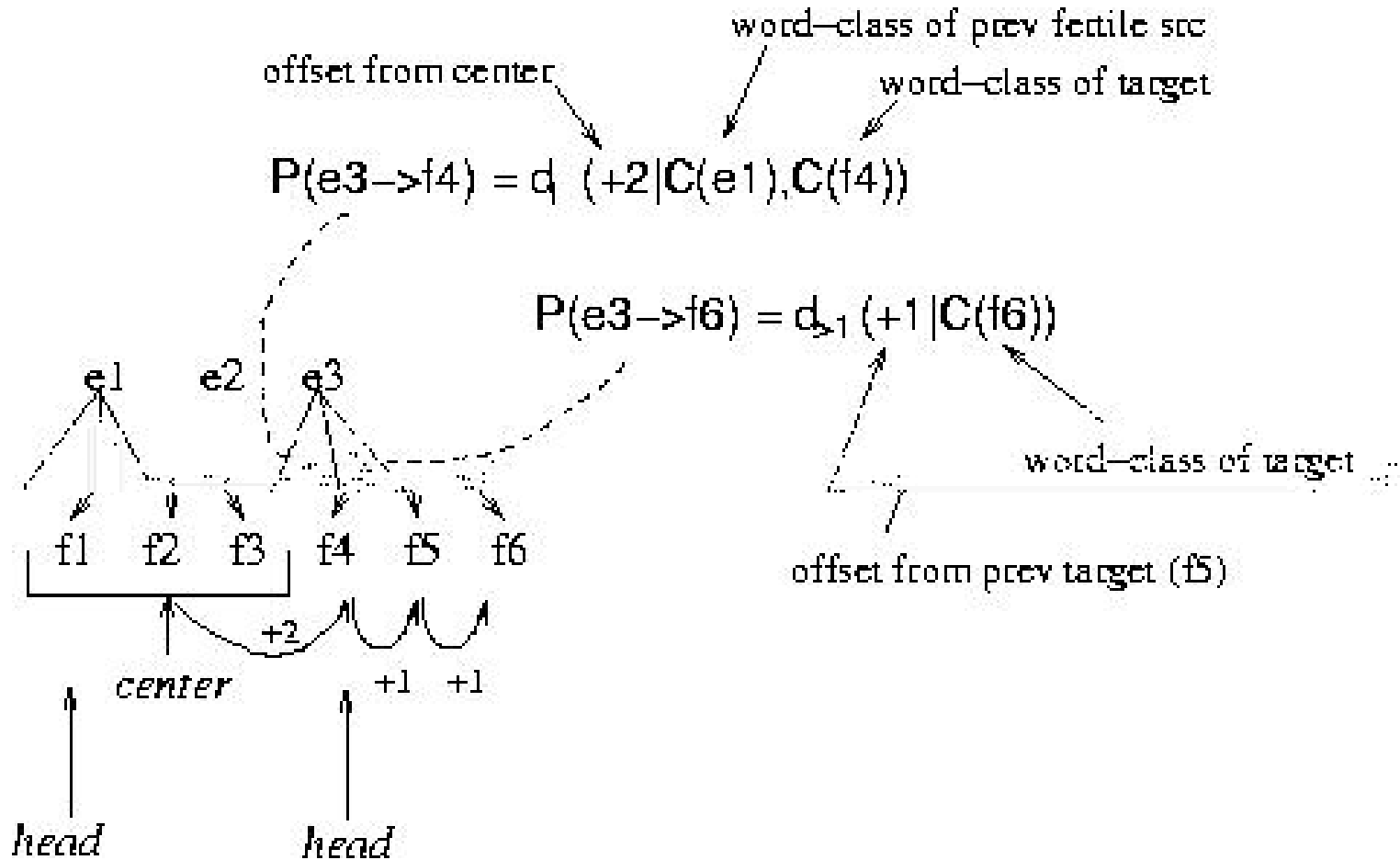
# IBM Model (word-based model)



# Bootstrapping IBM models

- Model 1: uniform distortion
  - Unique local maxima
  - Efficient EM algorithm (model 1-2)
- Model 2: general alignment:  $a(\text{epos}|\text{fpos}, \text{elen}, \text{flen})$
- Model 3: fertility:  $n(\phi|e)$ 
  - No full EM, count only neighbors (model 3-5)
  - Deficient (model 3-4)
- Model 4: relative distortion, word class
- Model 5: extra variables to avoid deficiency

# Model 4 distortion



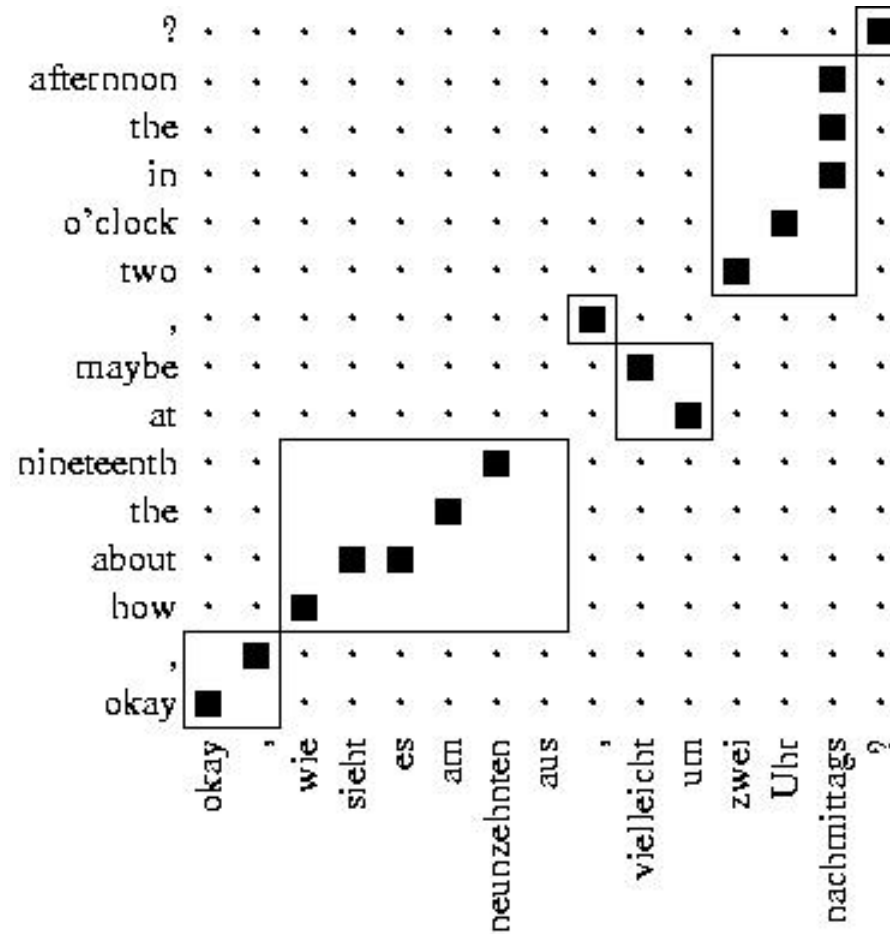
# Limitation of IBM models

- Only 1-to-N word mapping
- Handling fertility-zero words (difficult for decoding)
- Almost no syntactic information
  - Word class
  - Relative distortion
- Long-distance word movement

# Phrase-based models

- Wang's model [Wang and Waibel, 1998]
  - Use external phrase detector
  - Model 2 for phrase movement
  - Model 4 for within-phrase word movement
- Alignment Templates [Och et al., 1999]
  - Combine Model-4 alignments in two directions
  - Re-estimate phrase translation probabilities
  - Word class for templates

# Alignment Template



# Syntax-based models (1)

- Inversion Transduction Grammar [Wu, 1997]
  - Grammar formalism to generate two language
  - Binary branching only
  - Efficient dynamic programming decoder
- Head Automata [Alshawi et al., 2000]
  - Collection of two-head automata
  - Automatically induce bi-lingual dependency structure

# Inversion Transduction Grammar

[Wu, 1997]

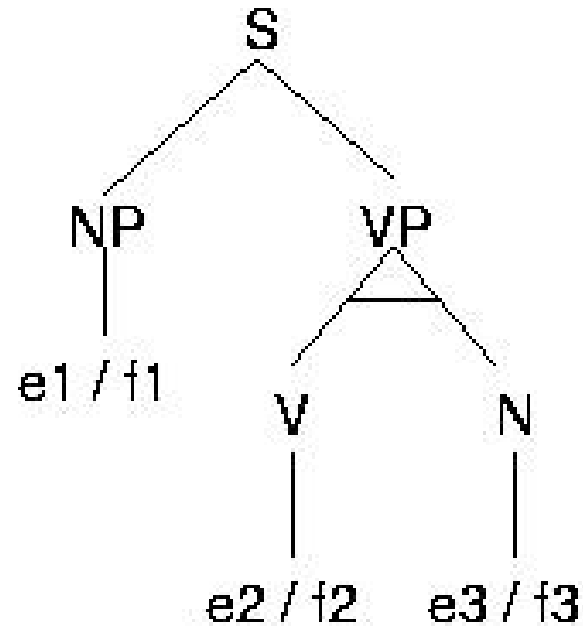
$S \rightarrow [NP VP]$

$NP \rightarrow e1 / f1$

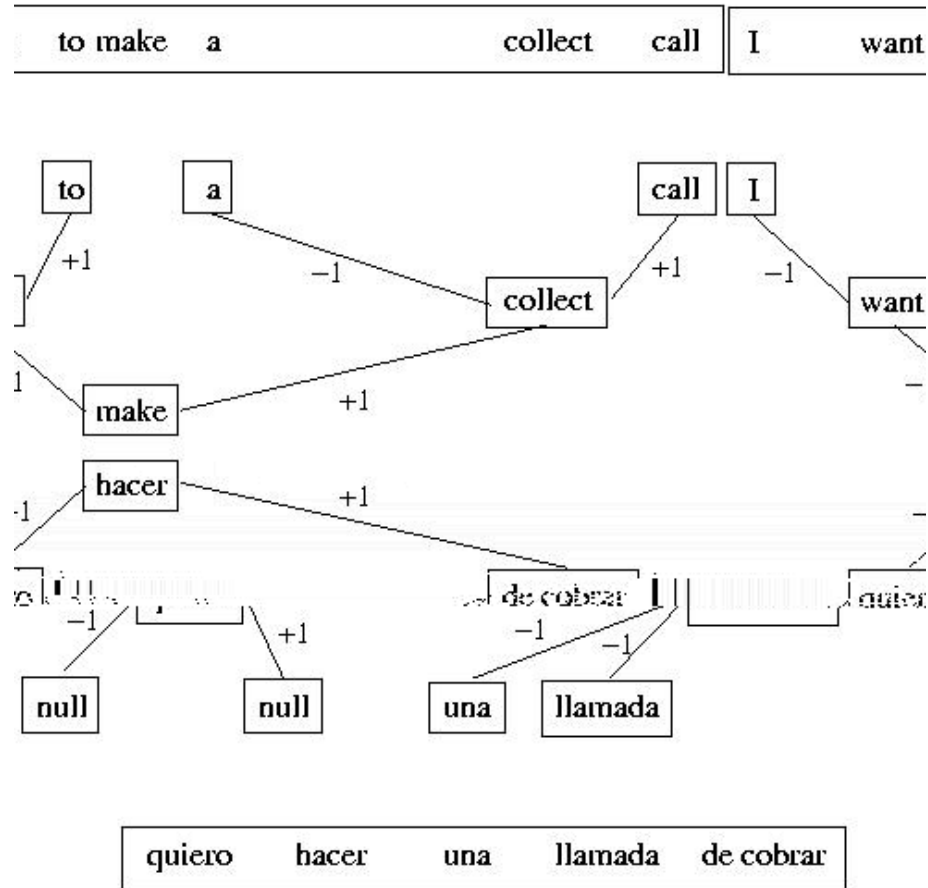
$VP \rightarrow \langle N V \rangle$

$V \rightarrow e2 / f2$

$N \rightarrow e3 / f3$



# Head Automata [Alshawi et al., 2000]



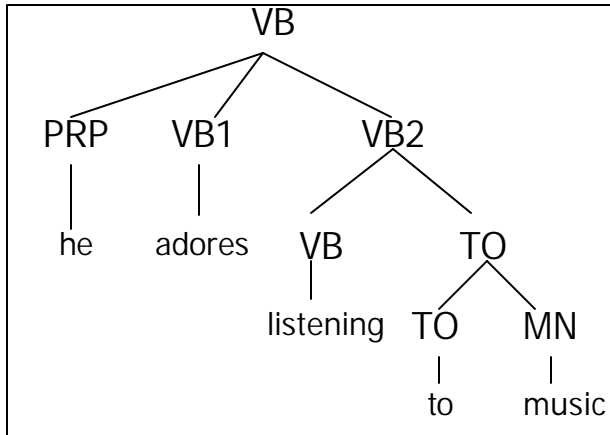
# Syntax-based models (2)

- Tree-to-string model [Yamada and Knight, 2001]
  - Use parser on one-side only
  - Insert, reorder, and translate operation
  - Efficient training (similar to inside-outside algorithm)
- Tree-to-tree models [Glidea 2003]
  - Subtree cloning operation

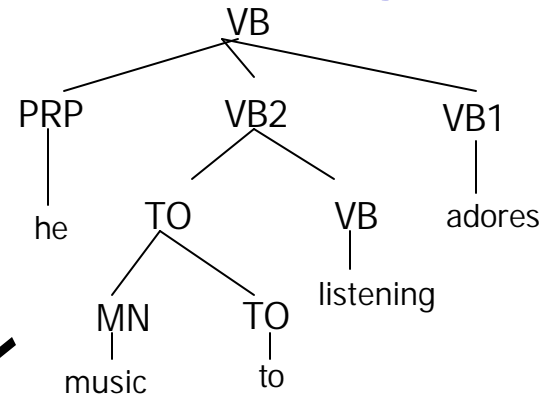
# Tree-to-string model

[Yamada and Knight, 2001]

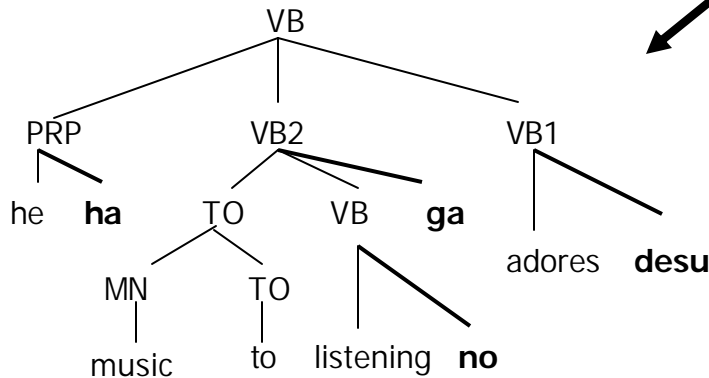
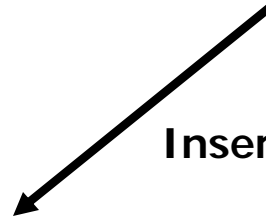
Parse Tree(E)



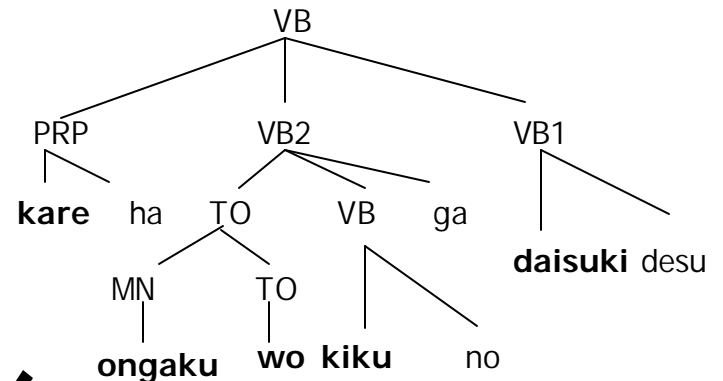
Reorder



Insert



Translate



Take Leaves



Sentence(J)

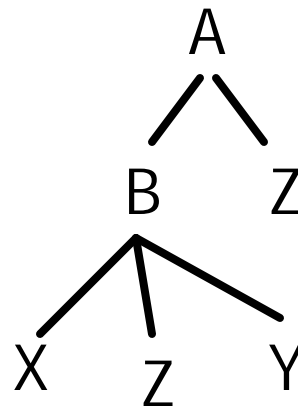
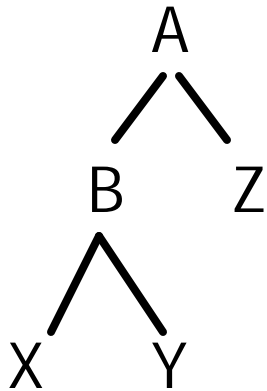
*Kare ha ongaku wo kiku no ga daisuki desu*

# Parameter Table: Reorder

Original Order	Reordering	P(reorder original)
<b>PRP VB1 VB2</b>	PRP VB1 VB2	0.074
	<b>PRP VB2 VB1</b>	<b>0.723</b>
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
<b>VB TO</b>	VB TO	0.107
	<b>TO VB</b>	<b>0.893</b>
<b>TO NN</b>	TO NN	0.251
	<b>NN TO</b>	<b>0.749</b>

# Tree-to-tree model [Glidea 2003]

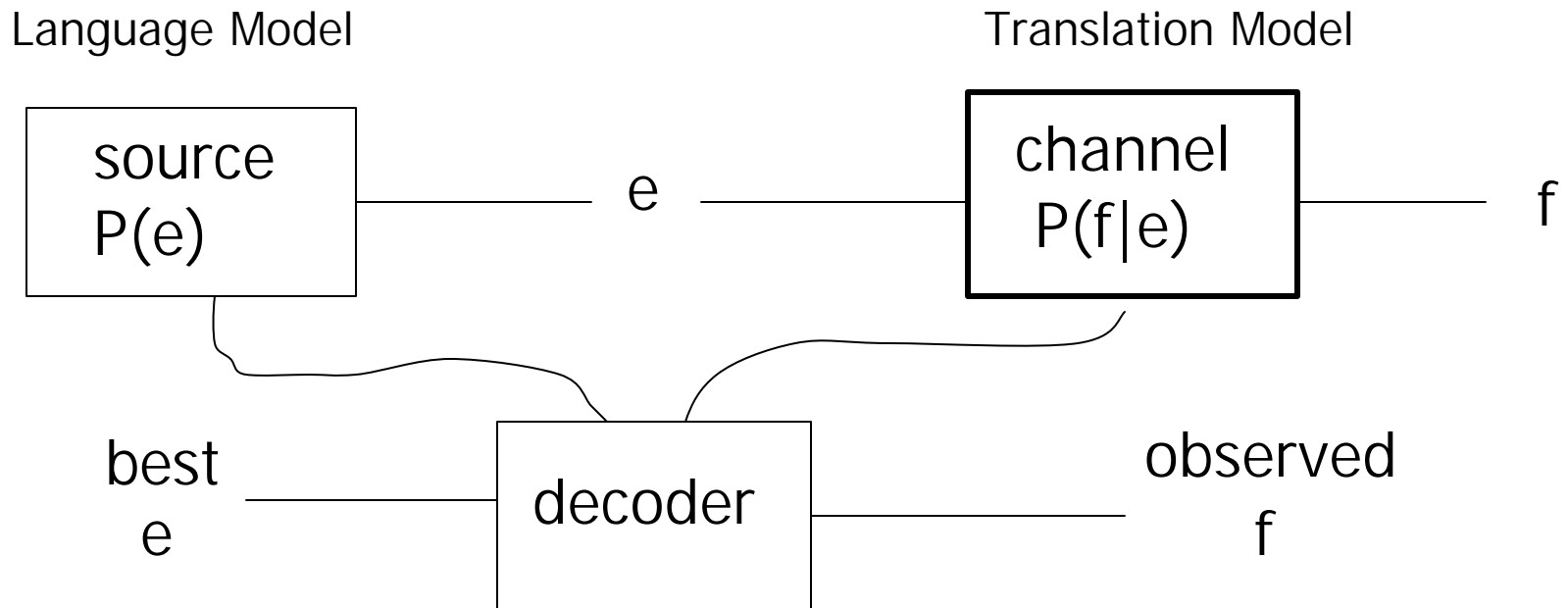
- Use parser on both sides
- Structural difference
- Subtree clone operation



# Other parts of the system

- Language Model
  - N-gram model
  - Syntax-based model
- Decoder
  - Heuristic search
  - Word-graph
  - Finite-state model

# Noisy Channel Model



$$\operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e)P(e)$$

# Summary

- Statistical MT
  - Data-driven approach
  - Model, parameter estimate, and search
- Simple models to complex models
  - Word-based
  - Phrase-based
  - Syntax-based