Phrase-Based MT

February 5, 2013

Translational Equivalence

*Ma hat die Prüfung **bestanden**, jedoch nur knapp*

Ma **insisted on** the test, but just barely.

Ma **passed** the test, but just barely.

How do lexical translation models deal with contextual information?

Translational Equivalence

*Ma hat die Prüfung **bestanden**, jedoch nur knapp*

Ma **insisted on** the test, but just barely.

Ma **passed** the test, but just barely.

F	E	
bestanden	**insisted**	-1.18
	were	-1.18
	existed	-1.36
	was	-1.39
	been	-1.43
	passed	-1.52
	consist	-1.87

Translational Equivalence

*Ma hat die Prüfung **bestanden**, jedoch nur knapp*

Ma **insisted on** the test, but just barely.

Ma **passed** the test, but just barely.

Lexical Translation

What is wrong with this?

How can we improve this?

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \overline{\mathbf{e}}, \overline{\mathbf{f}} \rangle \in \mathbf{a}} p(\overline{\mathbf{f}} \mid \overline{\mathbf{e}})$$

$\mathbf{f} =$ `Morgen fliege ich nach Baltimore zur Konferenz`

$\mathbf{e} =$ `Tomorrow I will fly to the Konferenz in Baltimore`

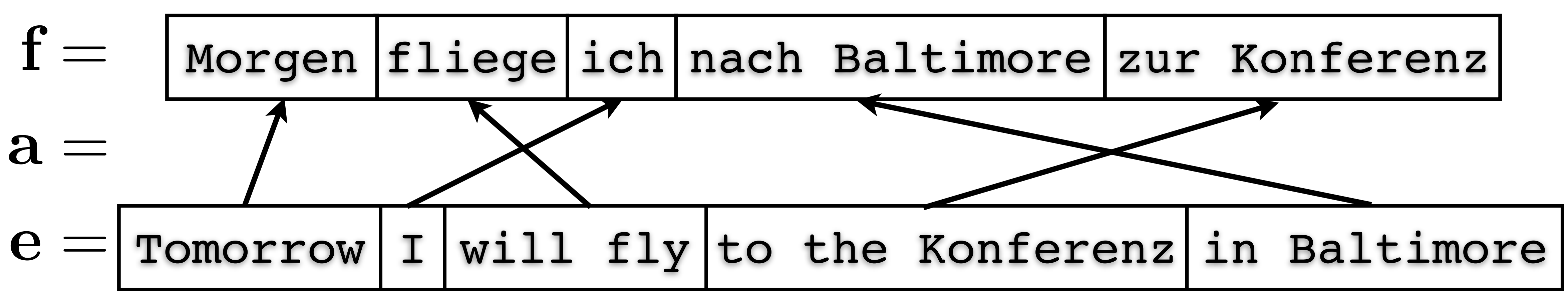
Translation Model

- What are the atomic units
  - Lexical translation: **words**
  - Phrase-based translation: **phrases**
- Benefits
  - many-to-many translation
  - use of local context in translation
- Downsides
  - Where do phrases comes from?
- Standard model used by Google, Microsoft ...

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

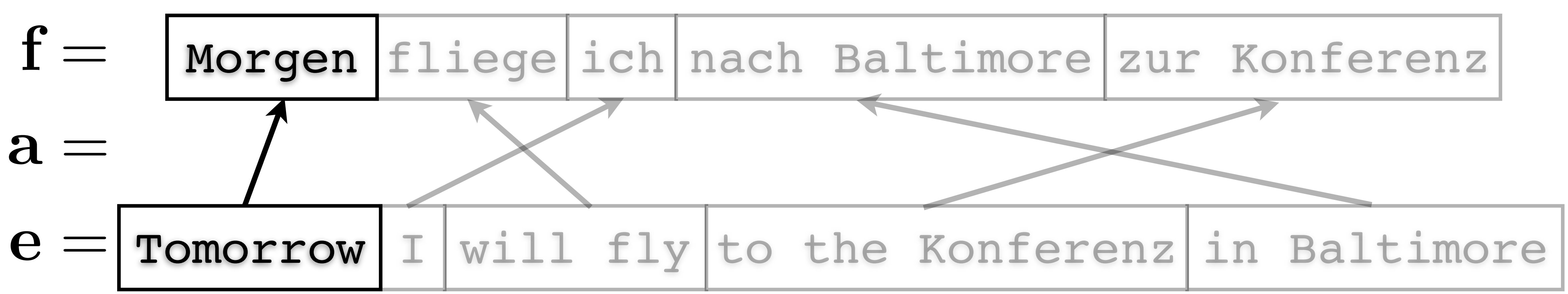
$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

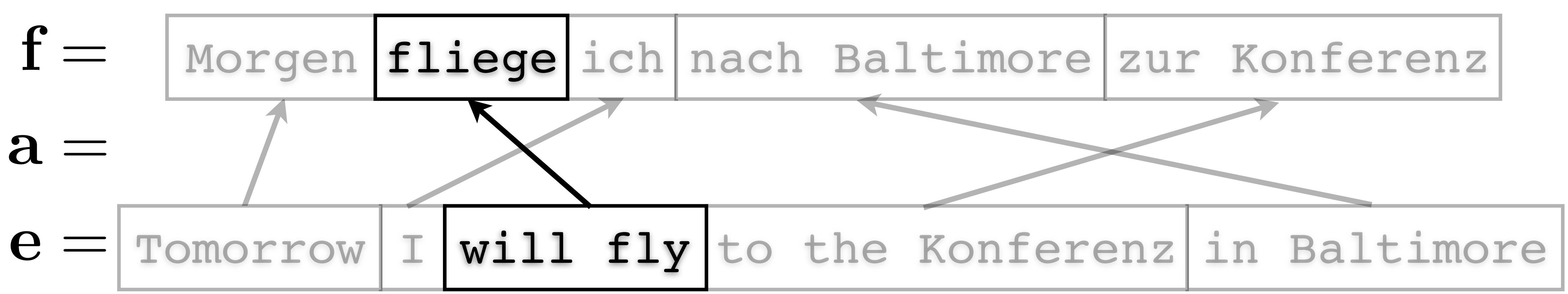


p(Morgen|Tomorrow)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

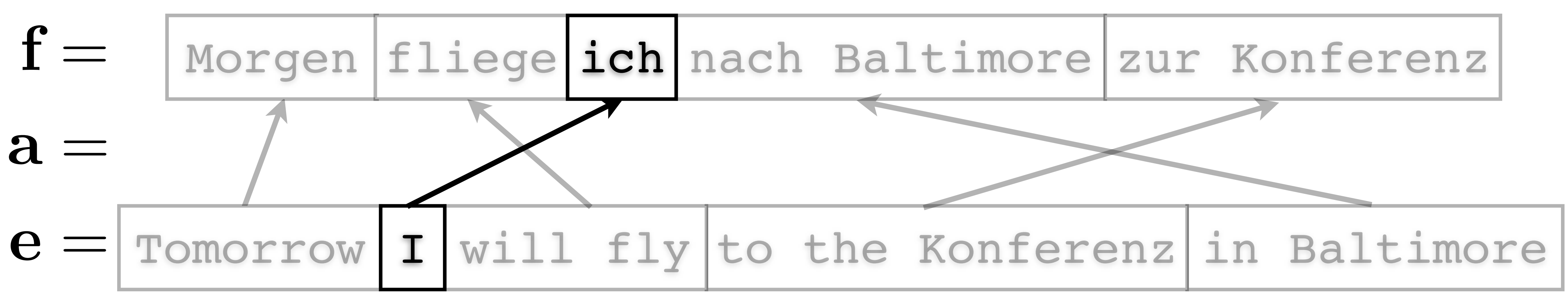


p(Morgen|Tomorrow) x *p*(fliege|will fly)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

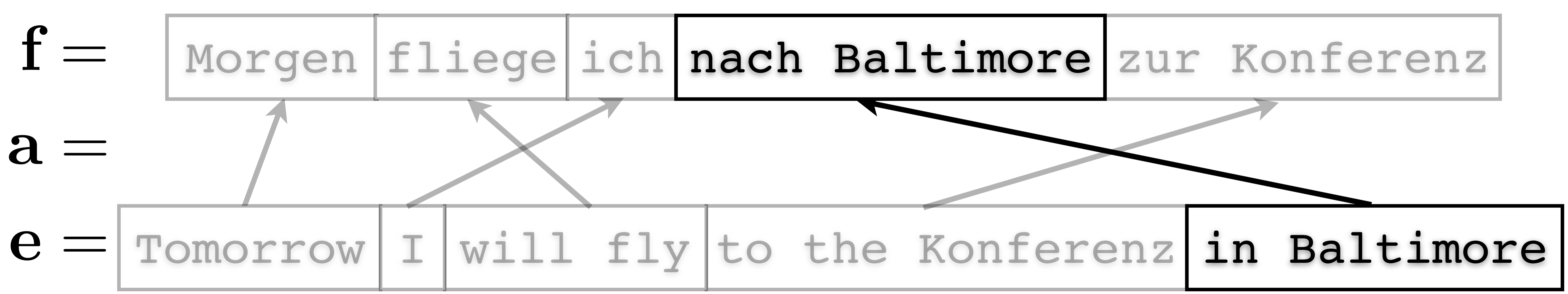


p(Morgen|Tomorrow) x *p*(fliege|will fly) x *p*(ich|I)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



p(Morgen|Tomorrow) x *p*(fliege|will fly) x *p*(ich|I) x ...

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

Marginalize to get p(f|e):

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

Phrases

- Contiguous strings of words
- Phrases are not necessarily syntactic constituents
- Usually have maximum limits
- Phrases subsume words (words are phrases)

Linguistic Phrases

- Model is not limited to linguistic phrases (NPs, VPs, PPs, CPs...)
- Non-constituent phrases are useful

  es gibt there is | there are

- Is a "good" phrase more likely to be
  [P NP] or [governor P]
  Why? How would you figure this out?

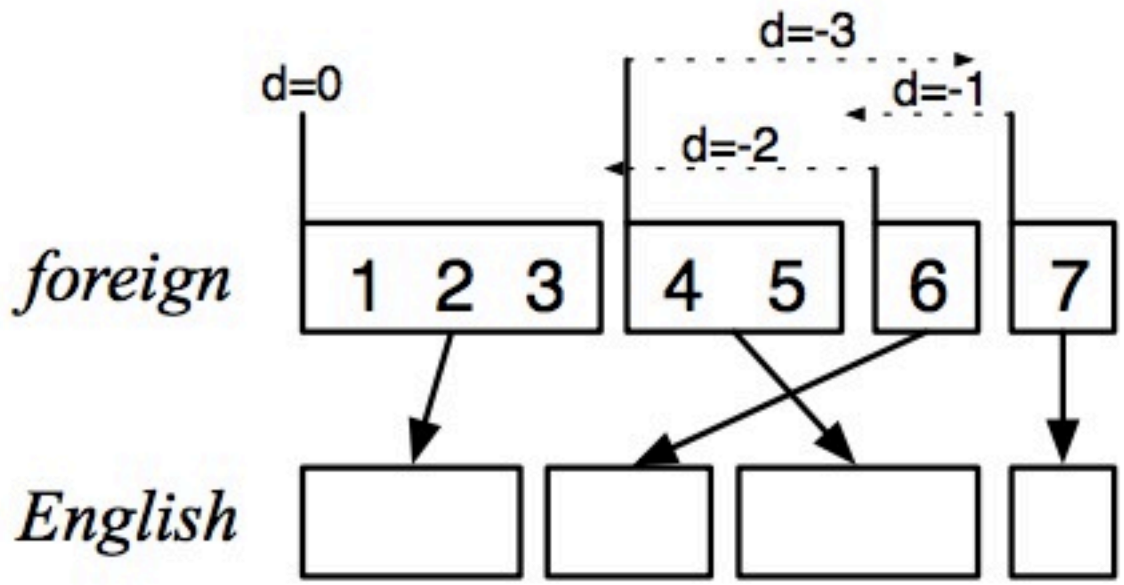
Phrase Tables

$\bar{\mathbf{f}}$	$\bar{\mathbf{e}}$	$p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$
das Thema	the issue	0.41
	the point	0.72
	the subject	0.47
	the thema	0.99
es gibt	there is	0.96
	there are	0.72
morgen	tomorrow	0.9
fliege ich	will I fly	0.63
	will fly	0.17
	I will fly	0.13

p(a)

- Two responsibilities
  - Divide the source sentence into phrases
    - Standard approach: uniform distribution over all possible segmentations
    - How many segmentations are there?
  - Reorder the phrases
    - Standard approach: Markov model on phrases (parameterized with log-linear model)

Reordering Model



phrase	translates	movement	distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	-3
4	7	skip over 6	+1

Scoring function: $d(x) = \alpha^{|x|}$ — exponential with distance

Learning Phrases

- Latent segmentation variable
- Latent phrasal inventory
- Parallel data
  - EM?

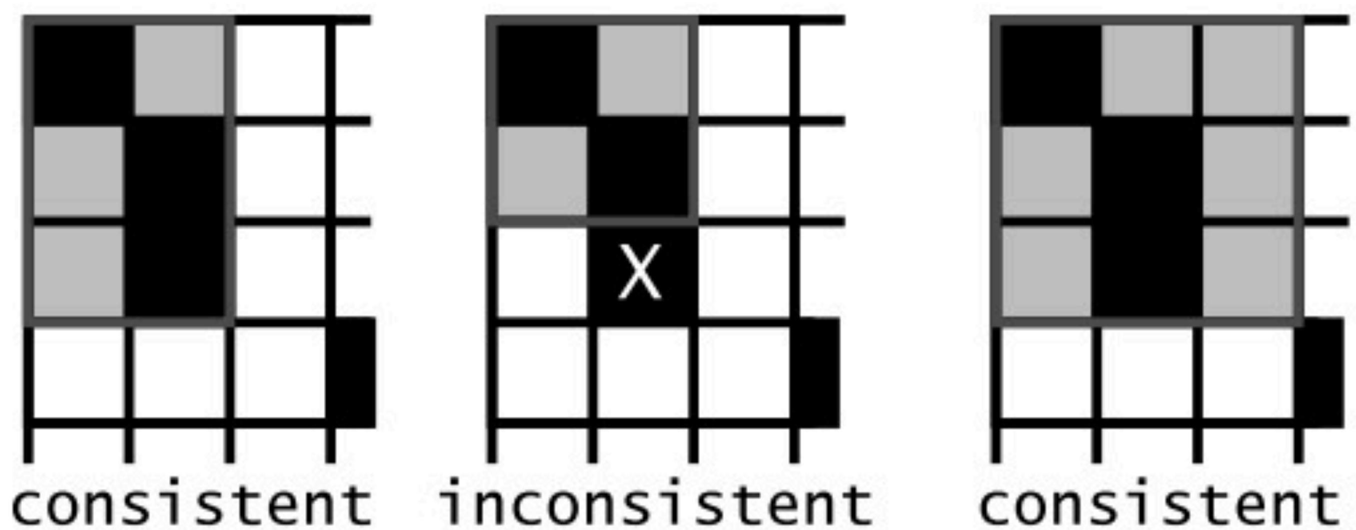
Computational problem: summing over all segmentations and alignments is #P-complete

Modeling problem: MLE has a degenerate solution.

Learning Phrases

- Three stages
  - word alignment
  - extraction of phrases
  - estimation of phrase probabilities

Consistent Phrases



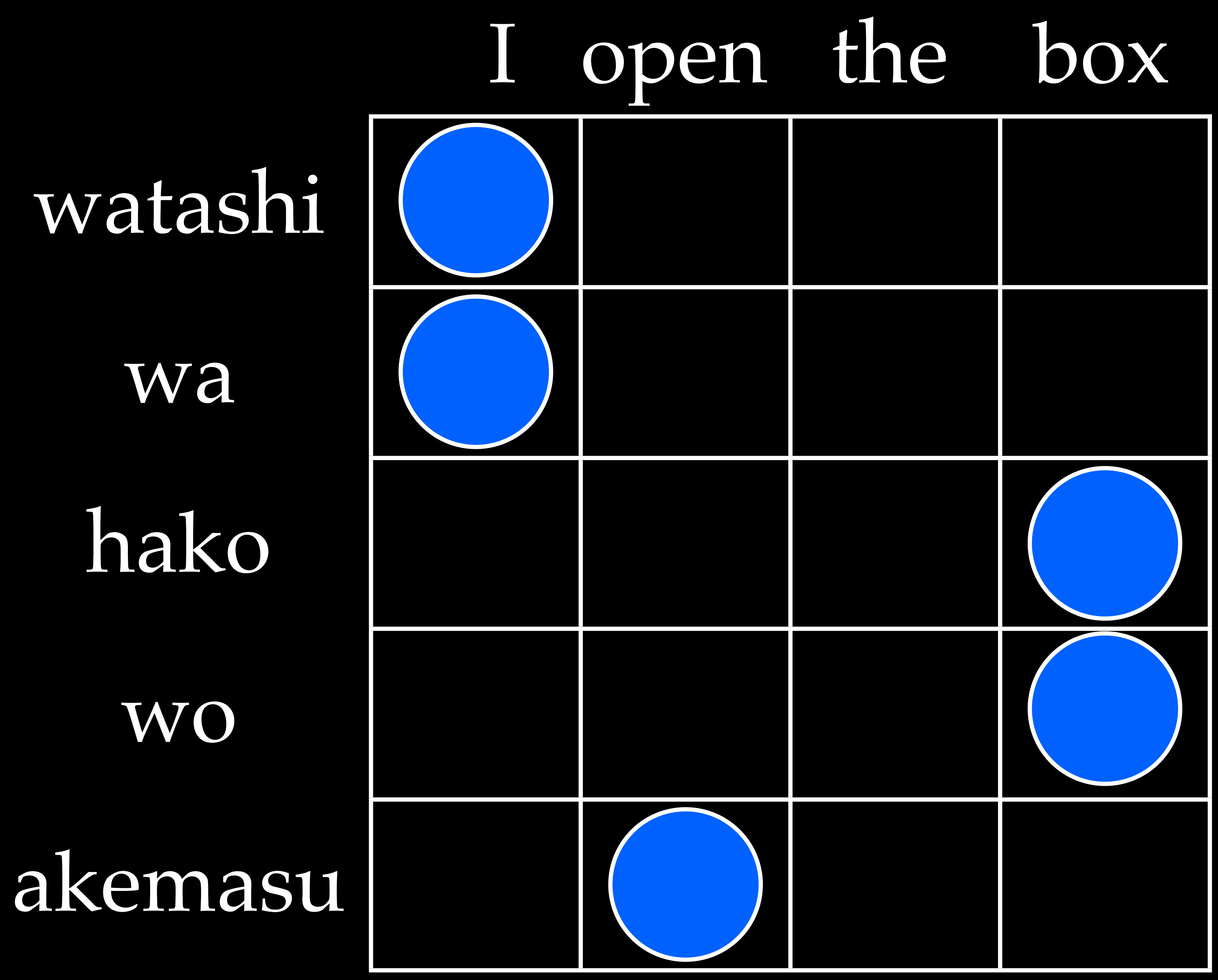
ok

violated
one alignment point outside

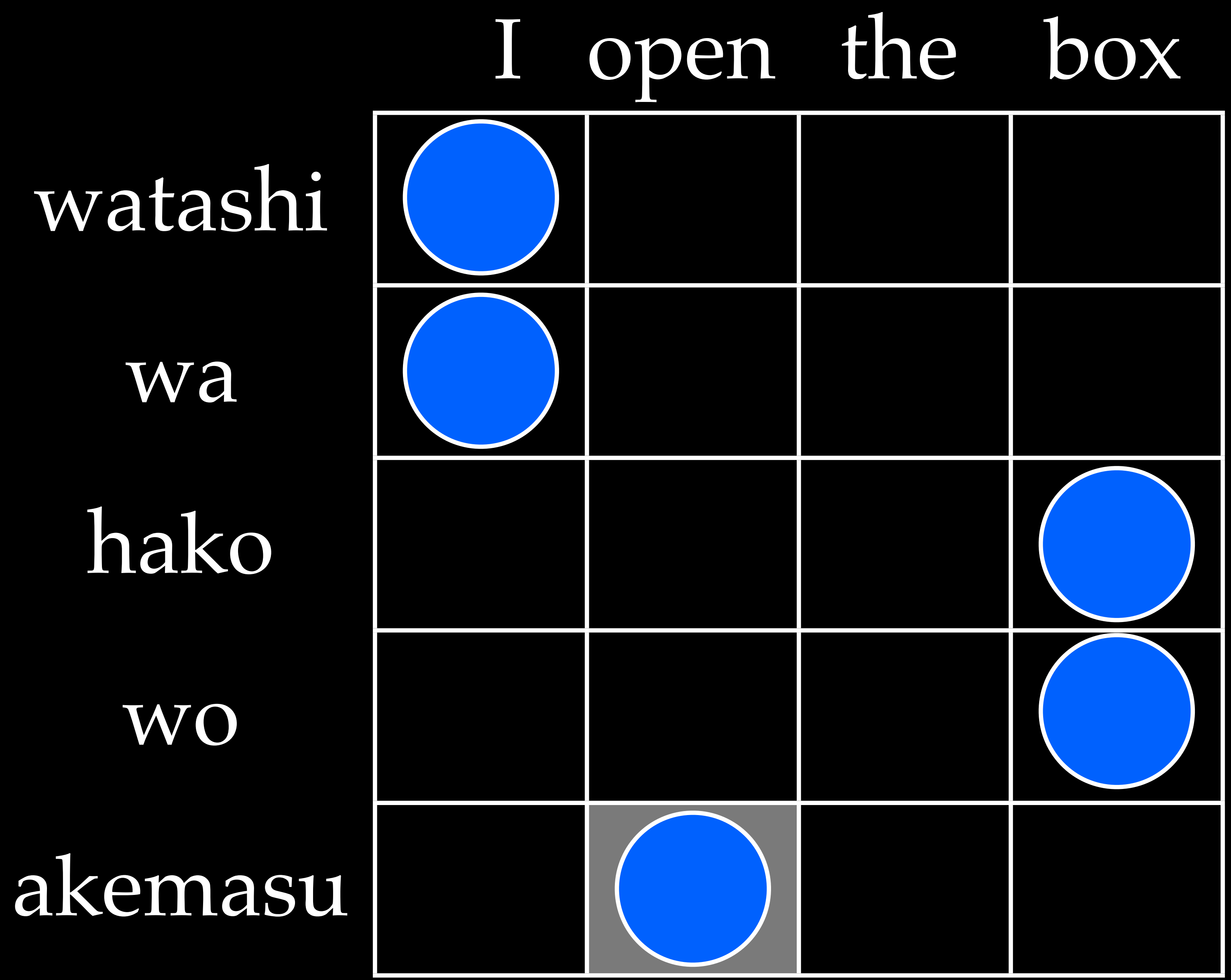
ok
unaligned word is fine

All words of the phrase pair have to align to each other.

Phrase Extraction

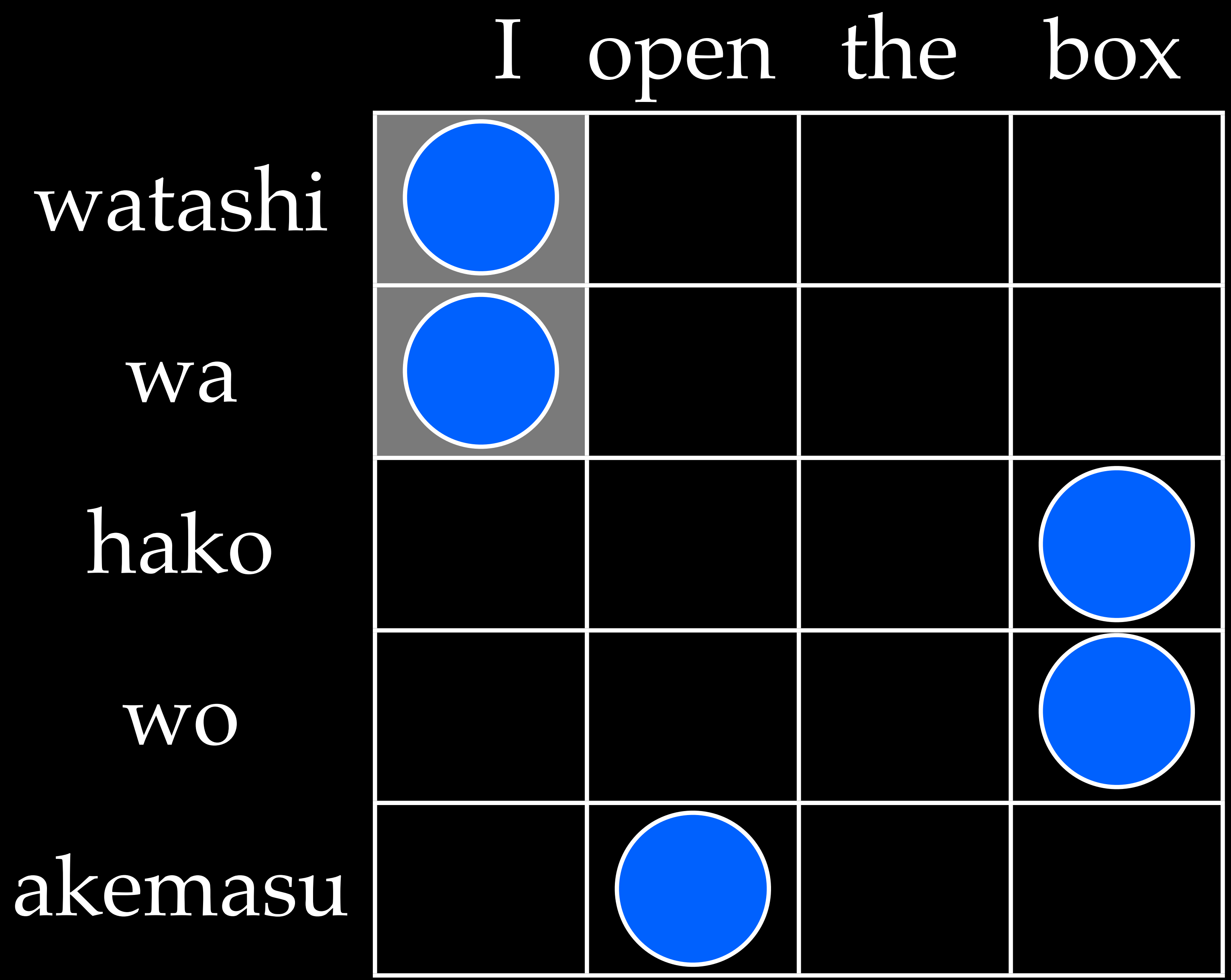


Phrase Extraction



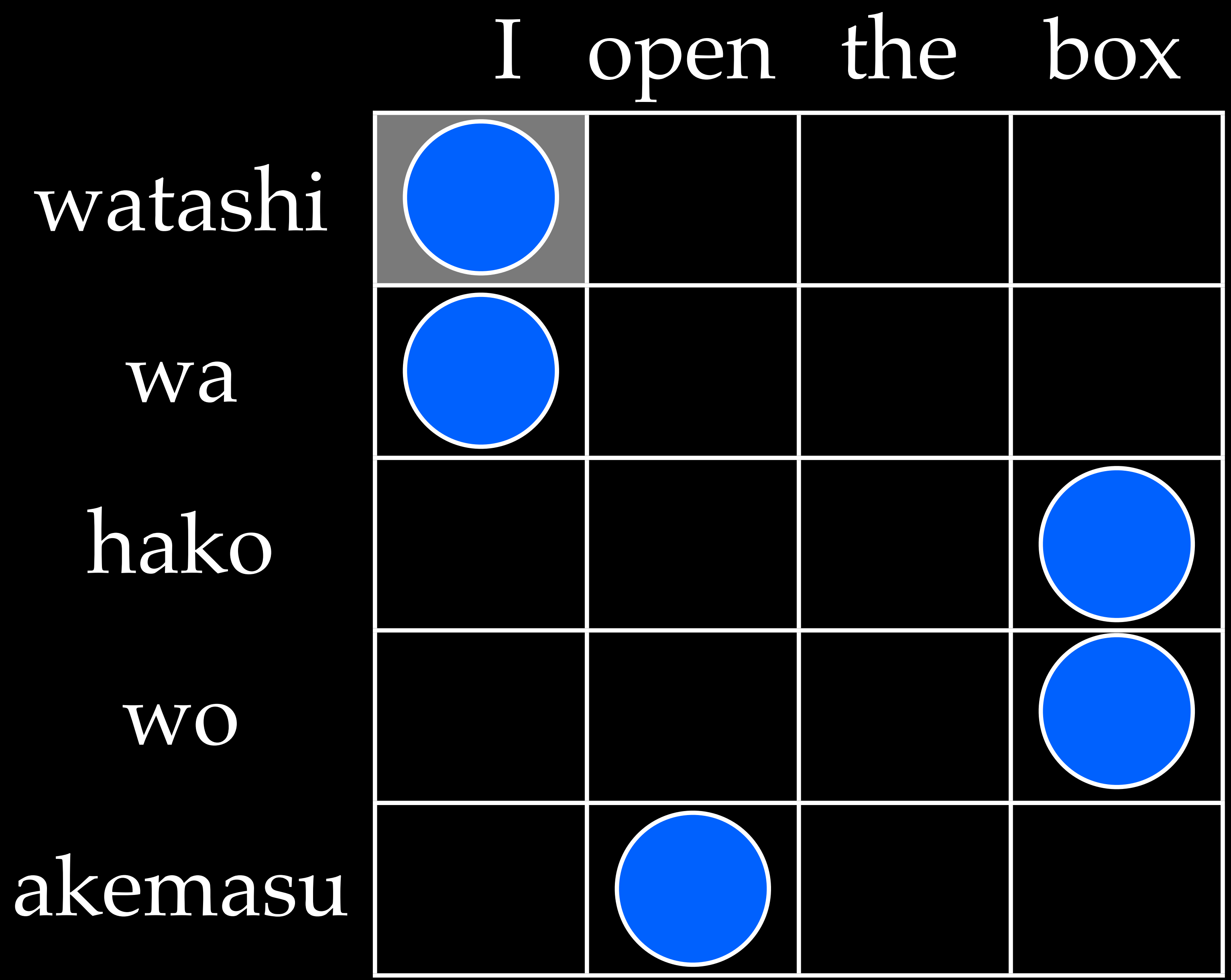
akemasu / open

Phrase Extraction



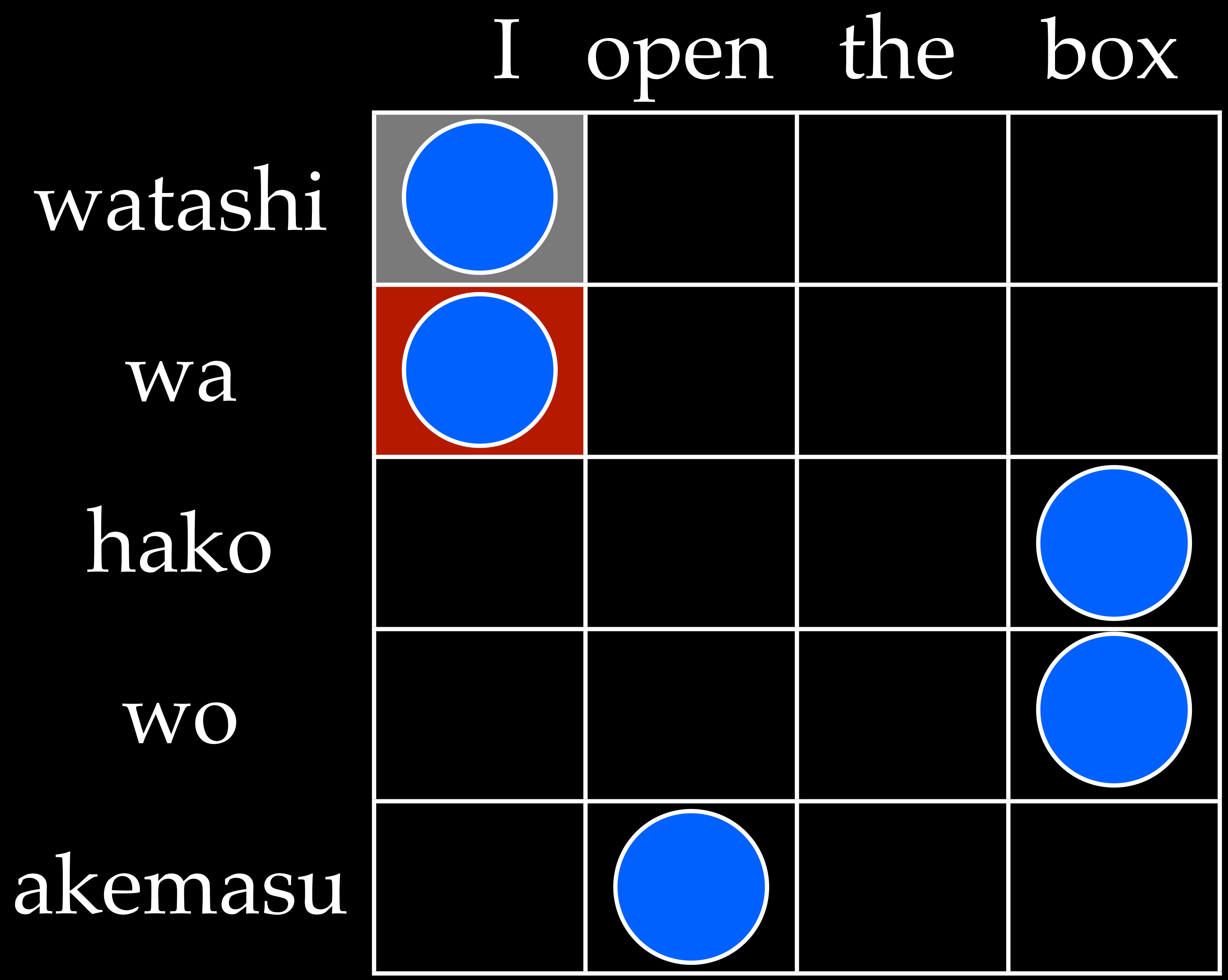
watashi wa / I

Phrase Extraction



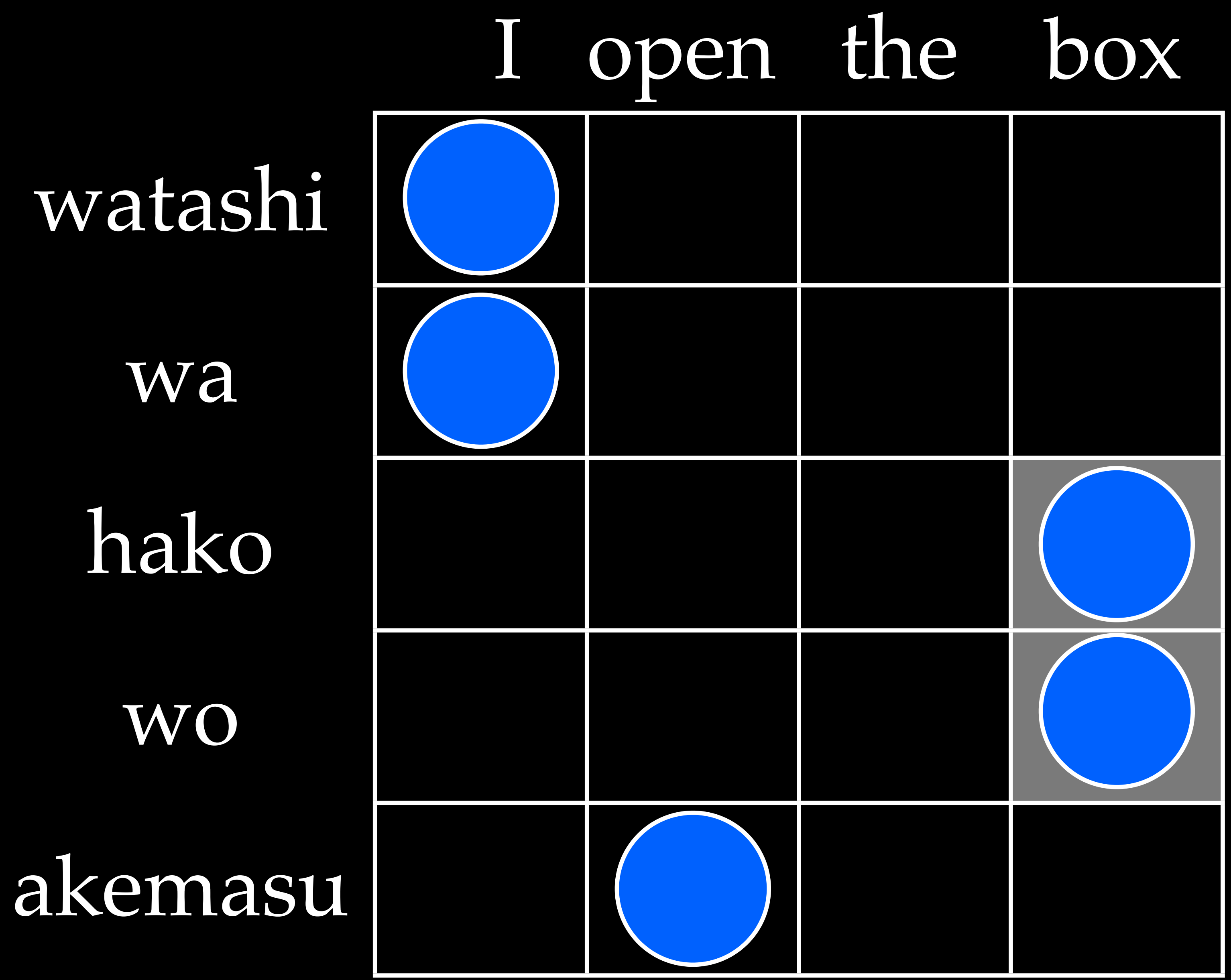
watashi / I

Phrase Extraction



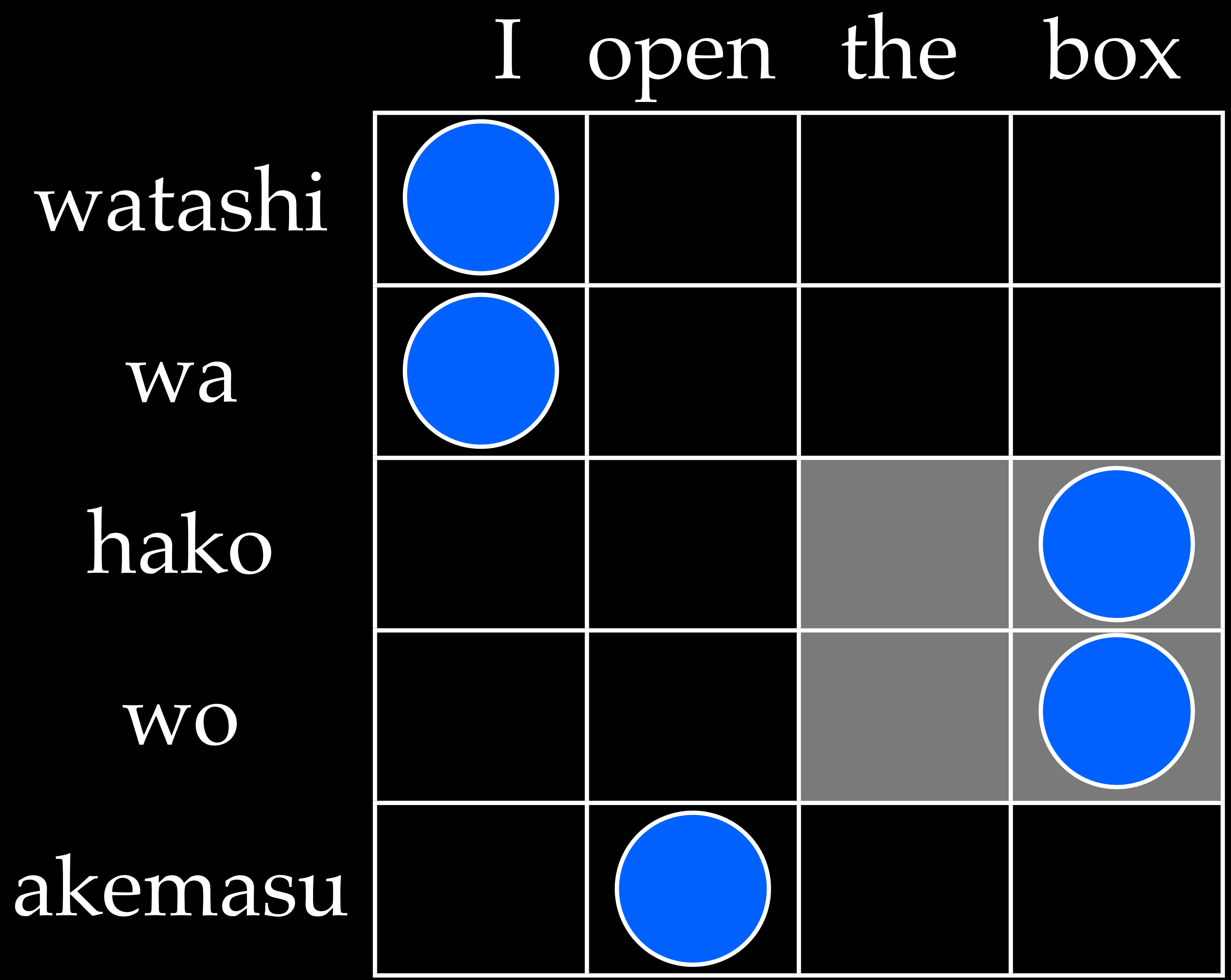
watashi wa / I

Phrase Extraction



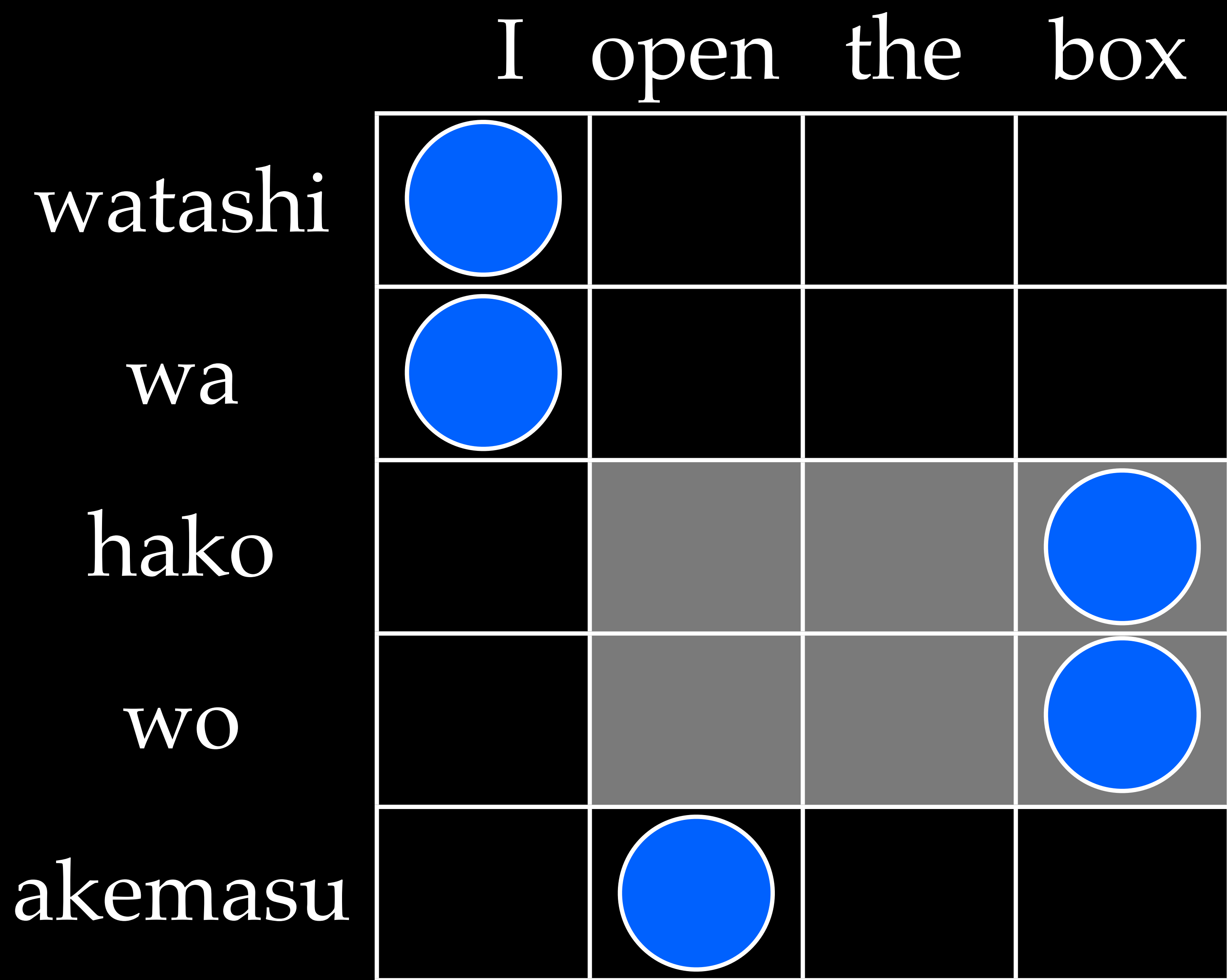
hako wo / box

Phrase Extraction



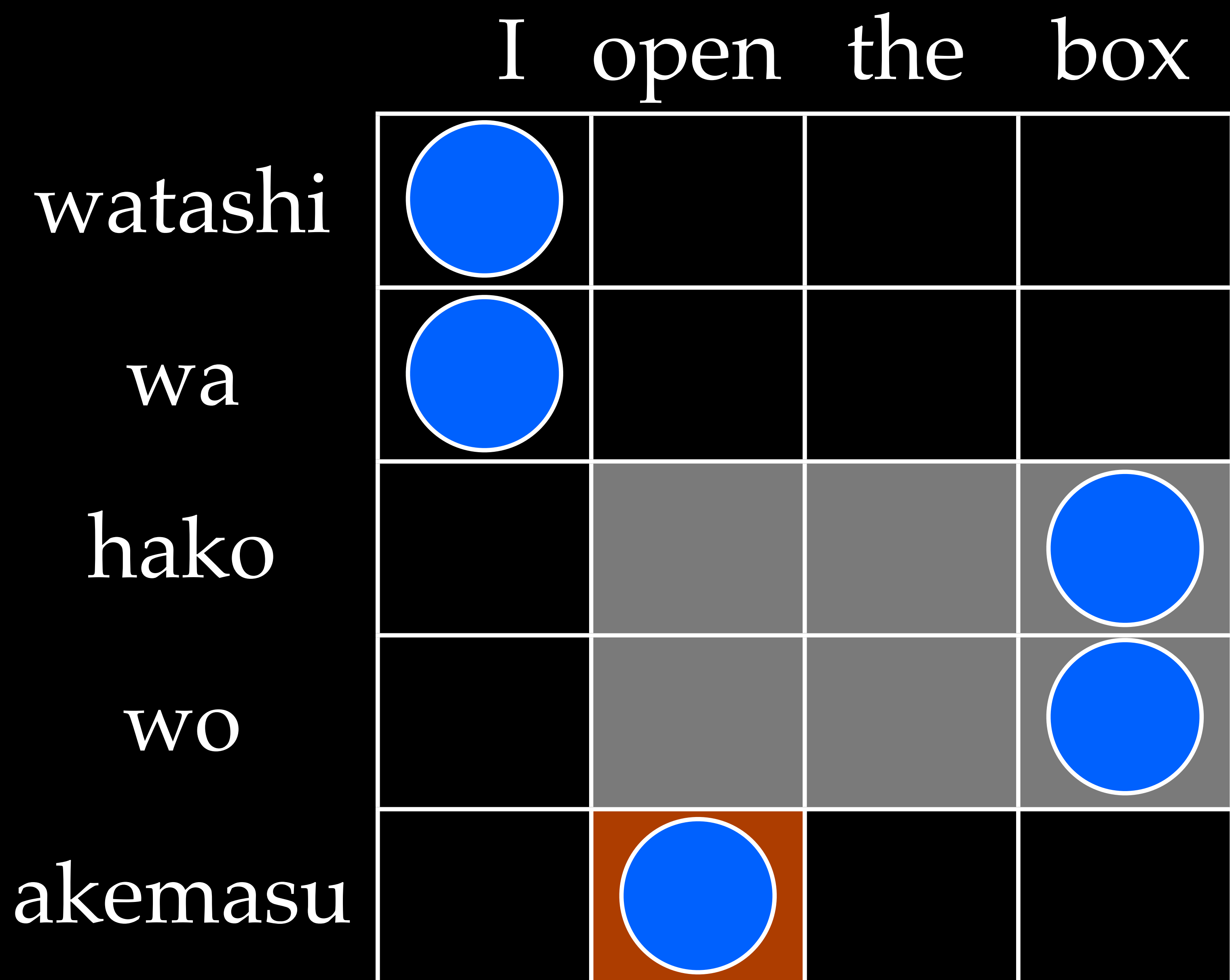
hako wo / the box

Phrase Extraction



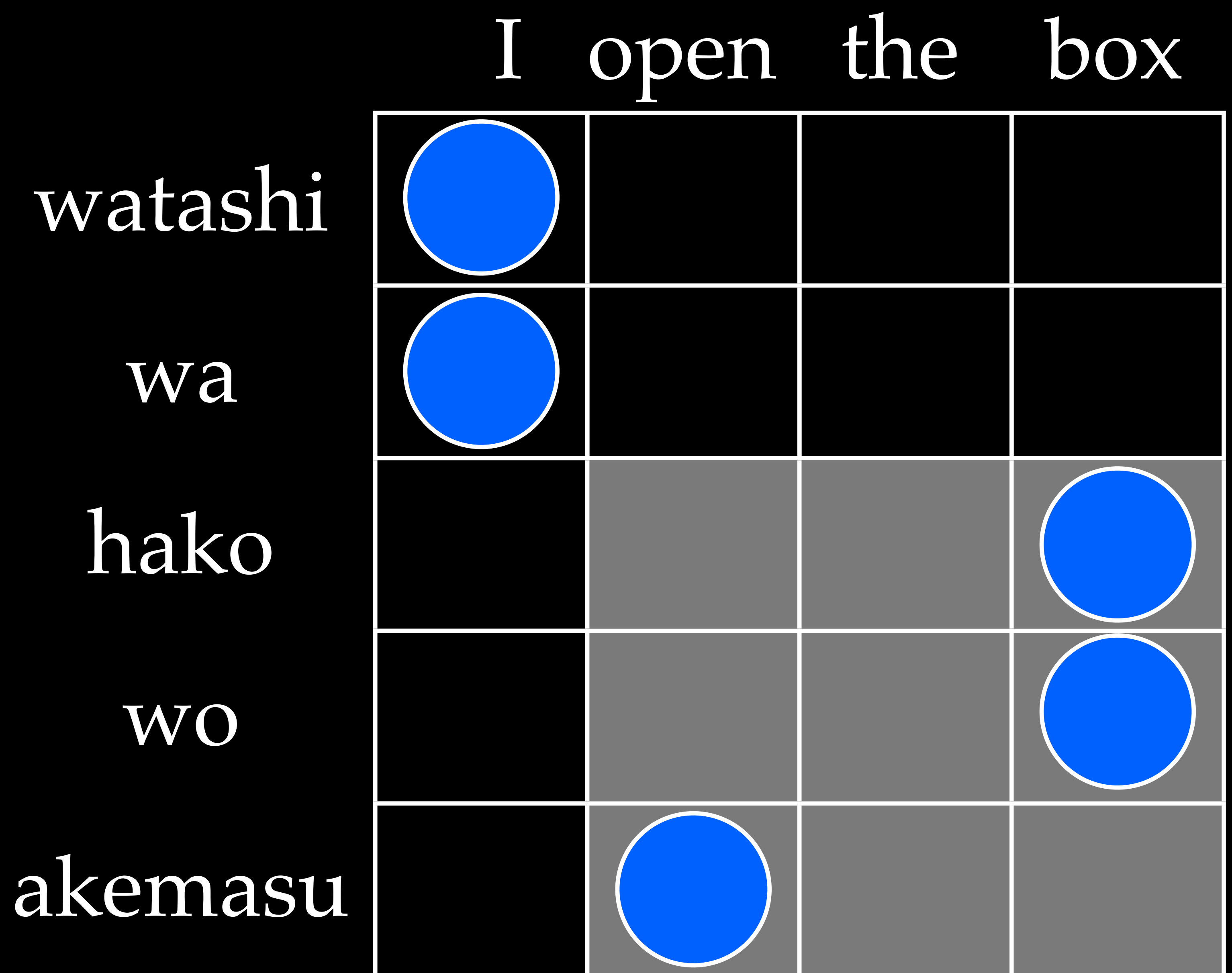
hako wo / open the box

Phrase Extraction



hako wo / open the box

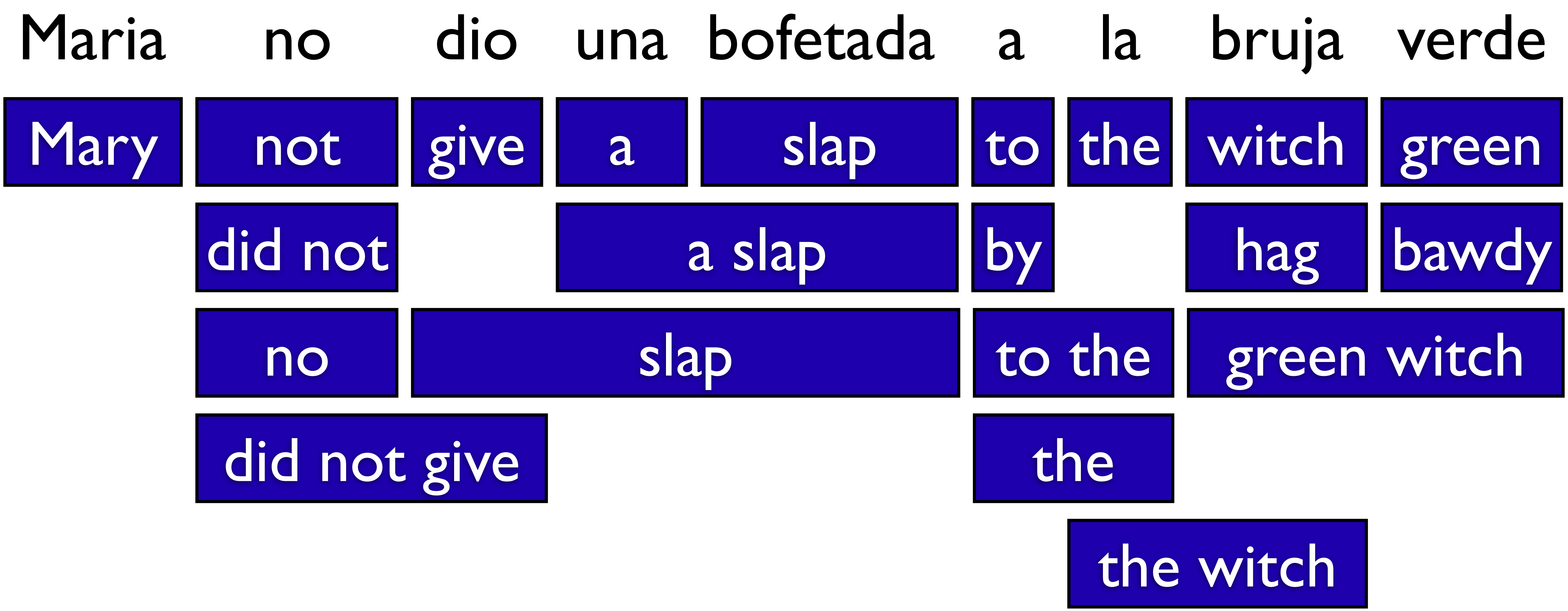
Phrase Extraction



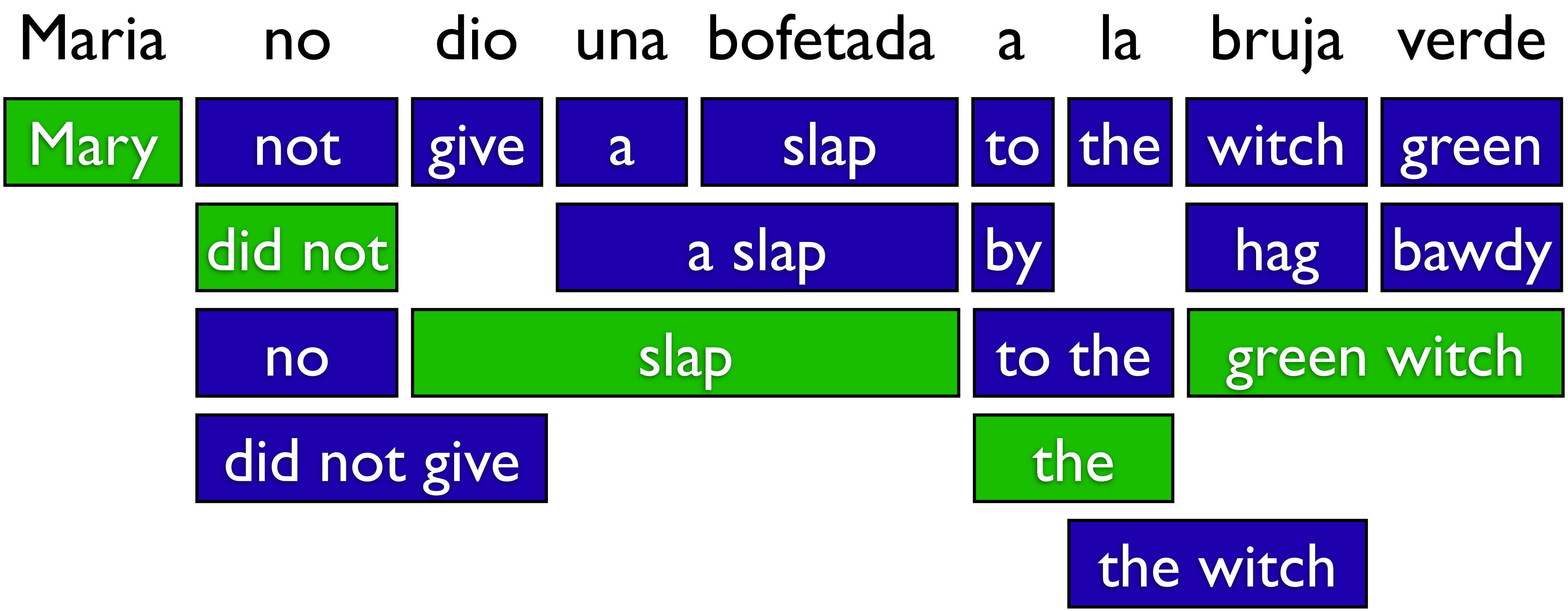
hako wo akemasu / open the box

Estimating Probabilities

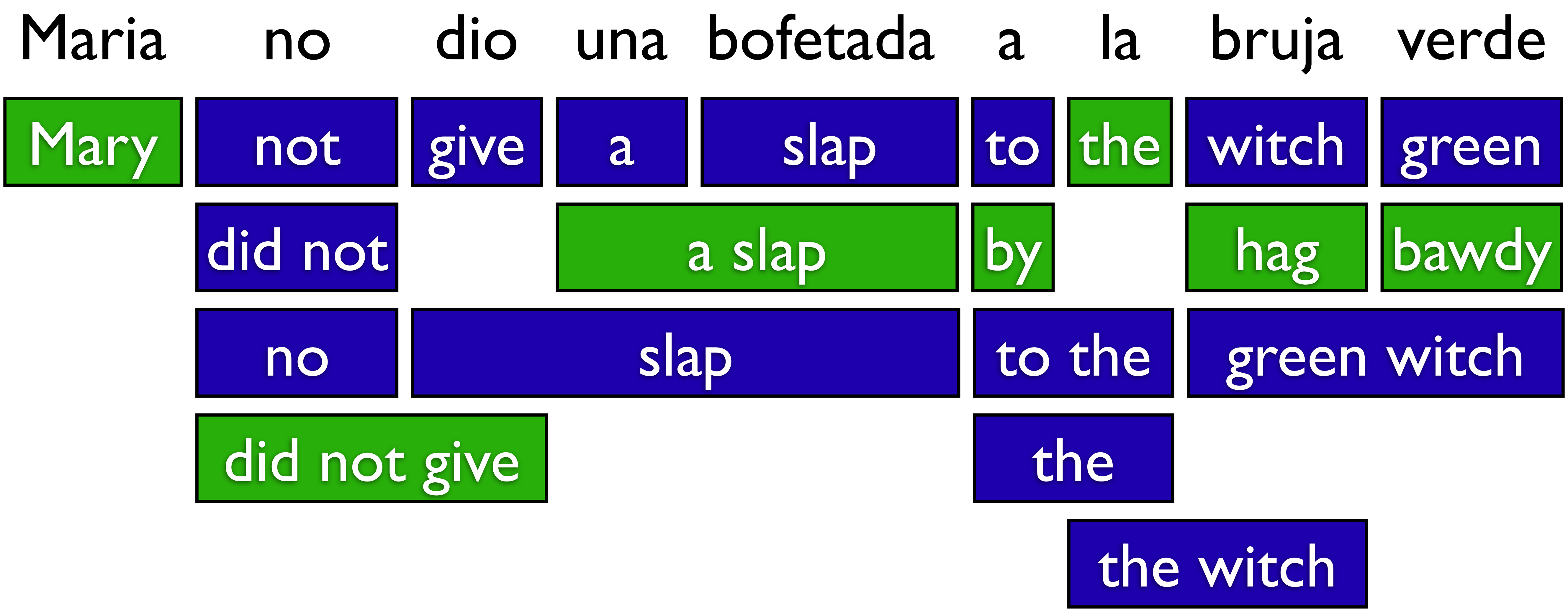
- What is the MLE?
  - Depends on the alignment model!
- Two options
  - EM over restricted space
  - Assume all alignments equally likely - count and normalize phrase pairs



Adapted from Koehn (2006)



Adapted from Koehn (2006)



Adapted from Koehn (2006)

Decoding algorithm

- Translation as a search problem
- Partial hypothesis keeps track of
  - which source words have been translated (*coverage vector*)
  - *n*-1 most recent words of English (for LM!)
  - a *back pointer* list to the previous hypothesis + (e,f) phrase pair used
  - the (partial) translation probability
  - the *estimated probability* of translating the remaining words (precomputed, a function of the coverage vector)
- **Start state**: no translated words, E=<s>, bp=nil
- **Goal state**: all translated words

Decoding algorithm

- Q[0] ← Start state
- for i = 0 to |**f**|-1
  - Keep *b* best hypotheses at Q[i]
  - for each hypothesis h in Q[i]
    - for each untranslated span in h.**c** for which there is a translation <e,f> in the phrase table
      - h' = h extend by <e,f>
      - Is there an item in Q[|h'.**c**|] with = LM state?
        - yes: update the item bp list and probability
        - no: Q[|h'.**c**|] ← h'
- Find the best hypothesis in Q[|**f**|], reconstruction translation by following back pointers

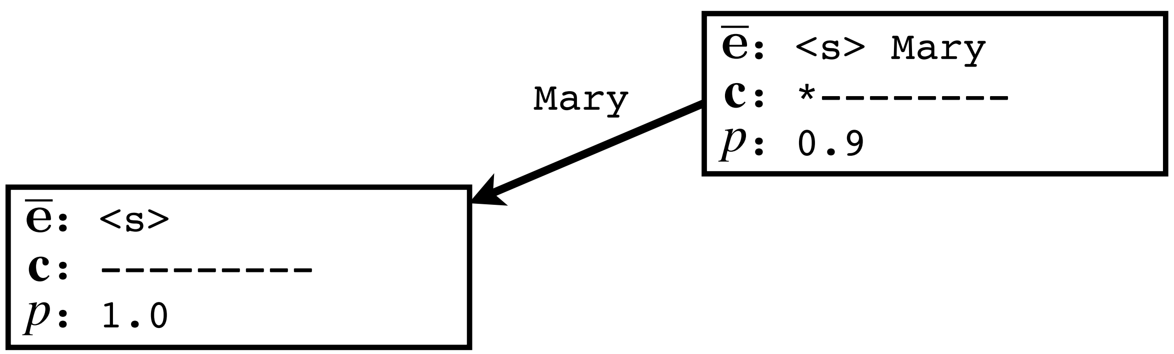
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...

$\bar{\mathbf{e}}$:	<s>
$\mathbf{c}$:	---------
p:	1.0

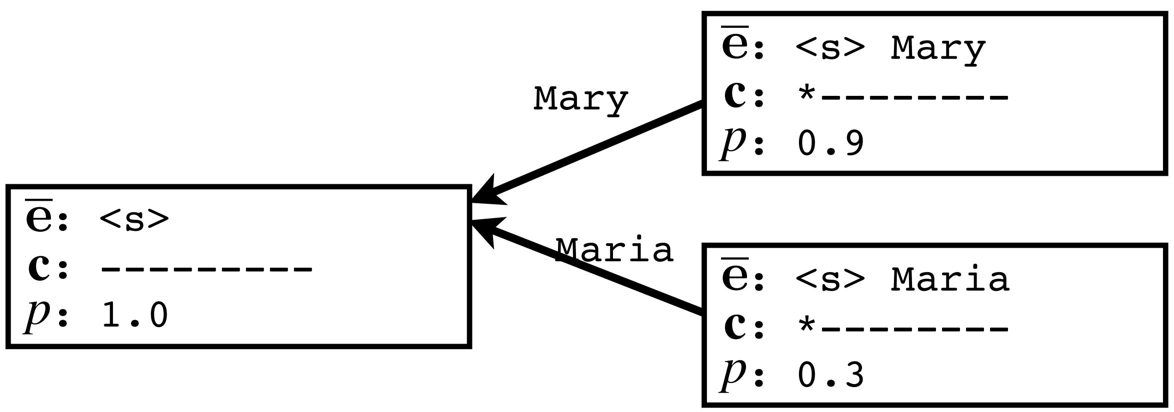
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



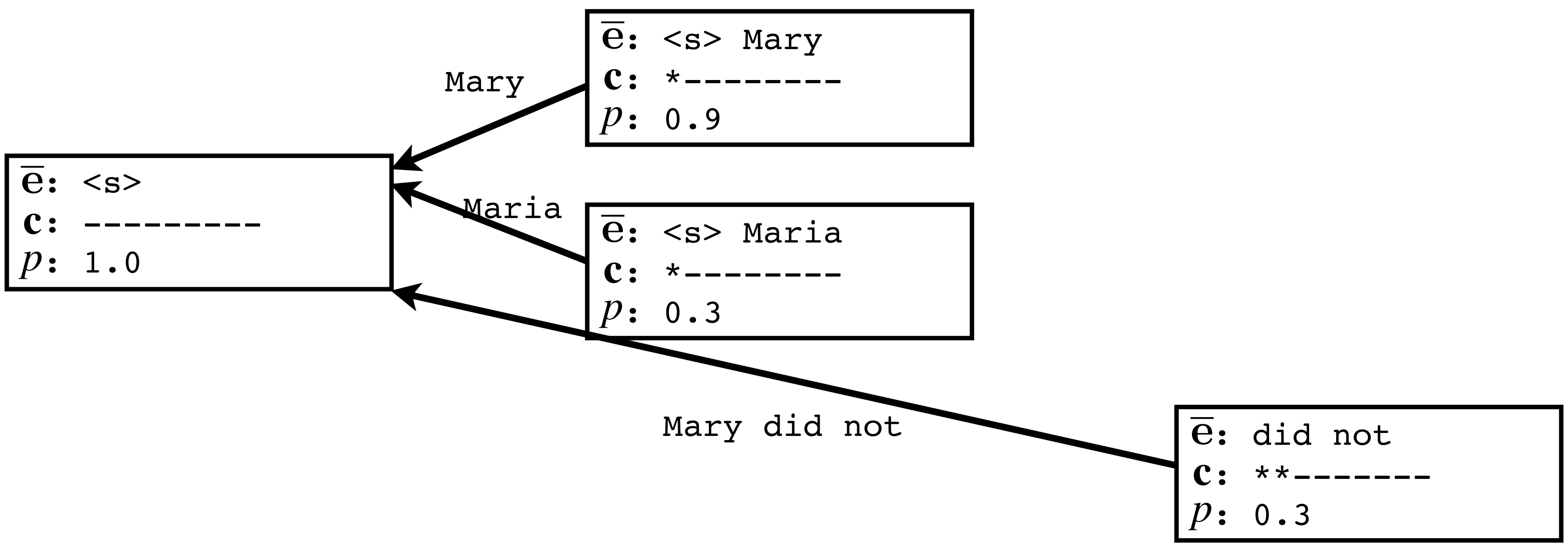
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



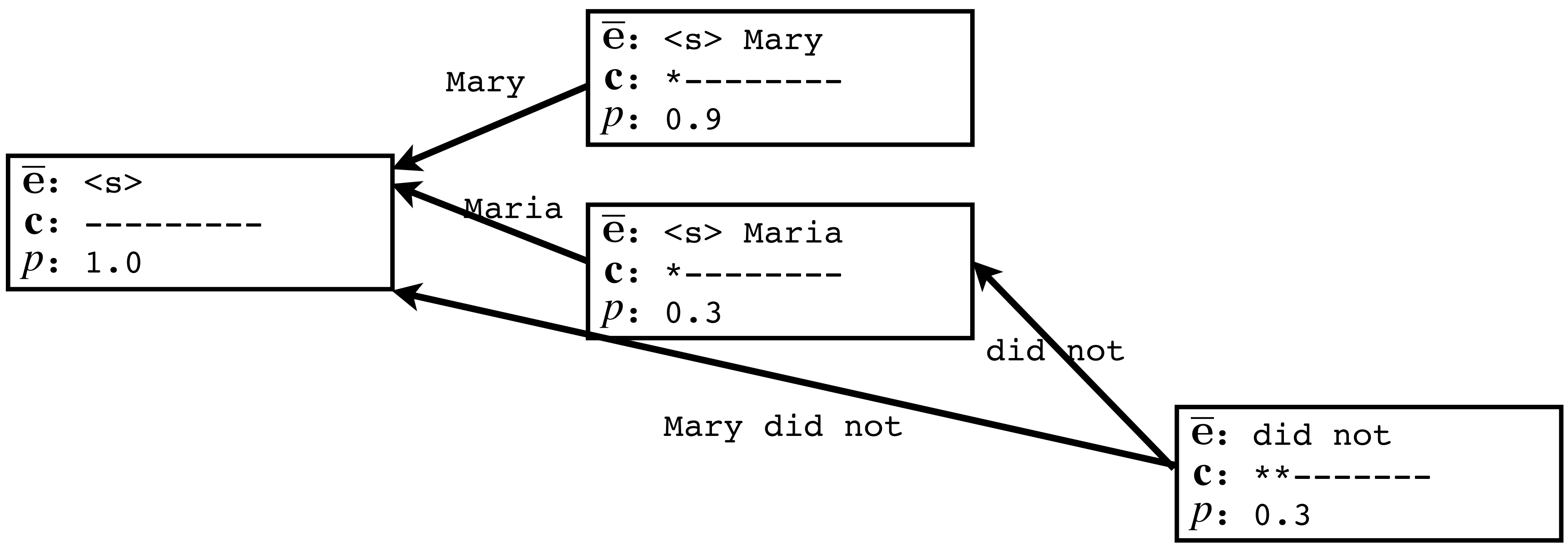
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



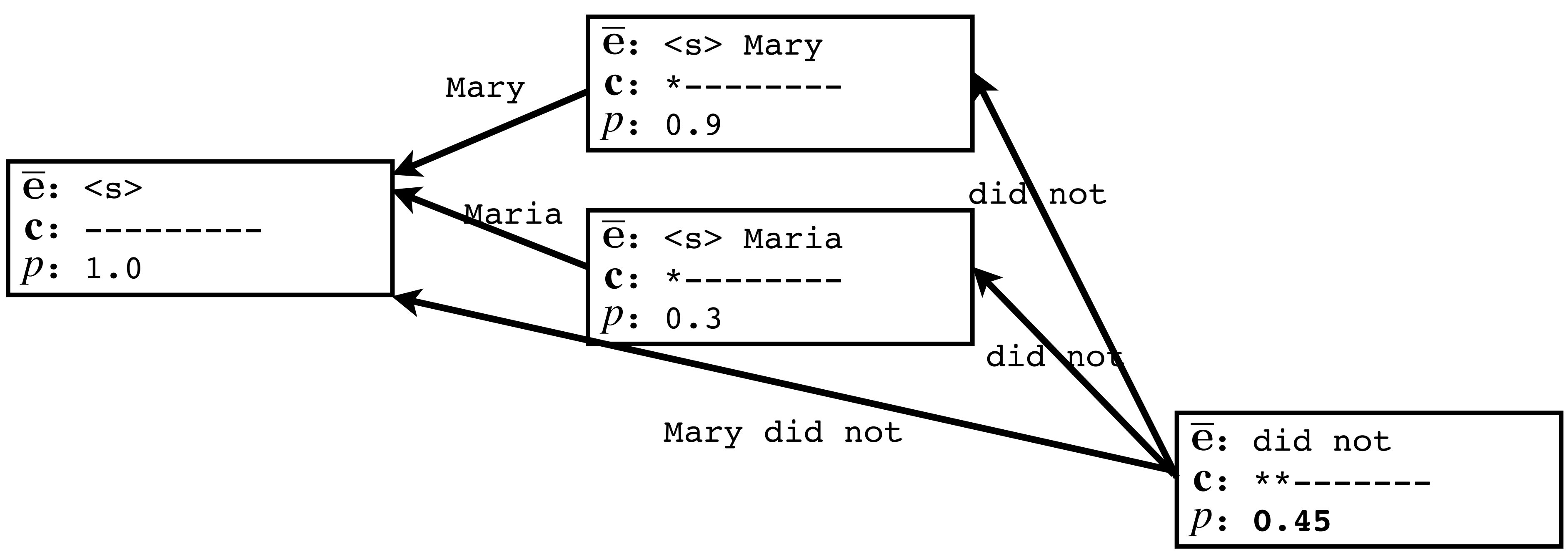
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



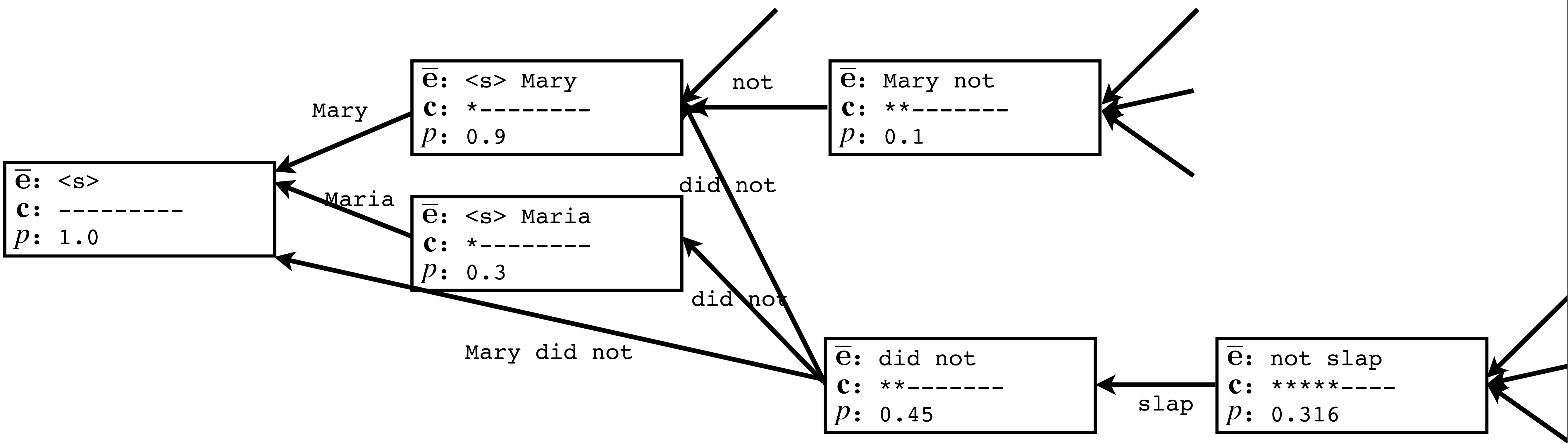
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...

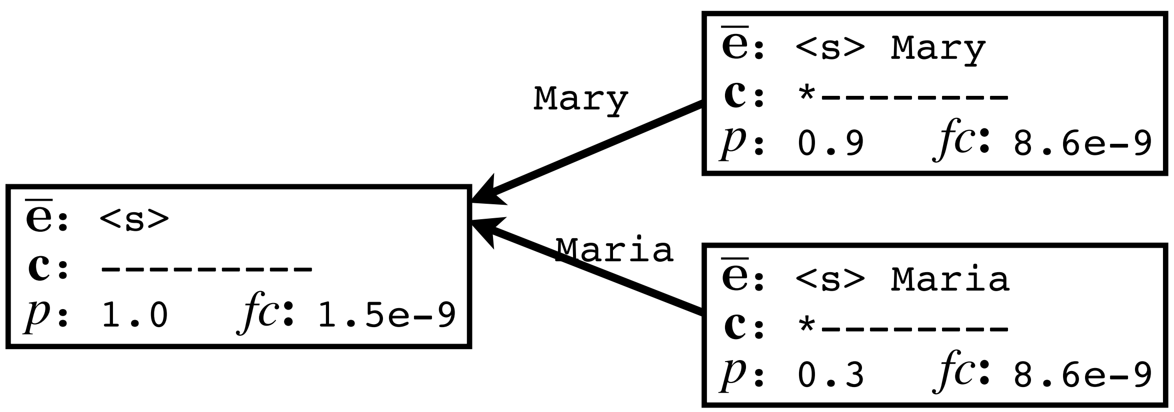


Reordering

- Language express words in different orders
  - bruja verde *vs.* green witch
- Phrase pairs can "memorize" some of these
- More general: in decoding, "skip ahead"
- Problem:
  - Won't "easy parts" of the sentence be translated first?
- Solution:
  - **Future cost estimate**
  - For every **coverage vector**, estimate what it will cost to translate the remaining untranslated words
  - When pruning, use p * future cost!

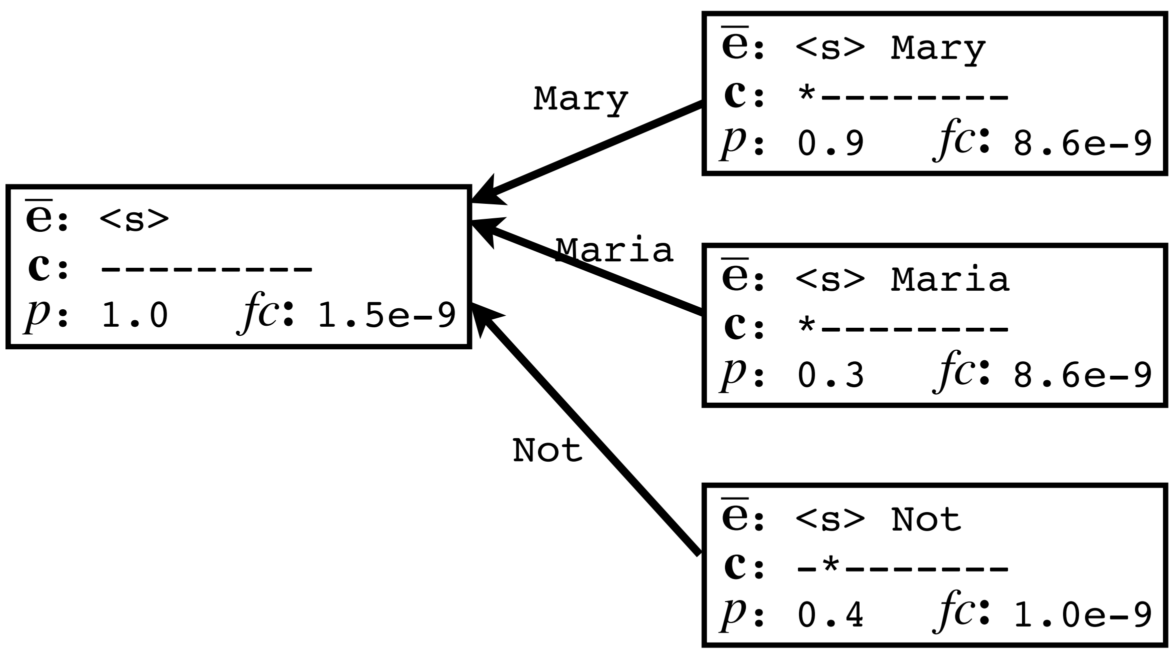
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



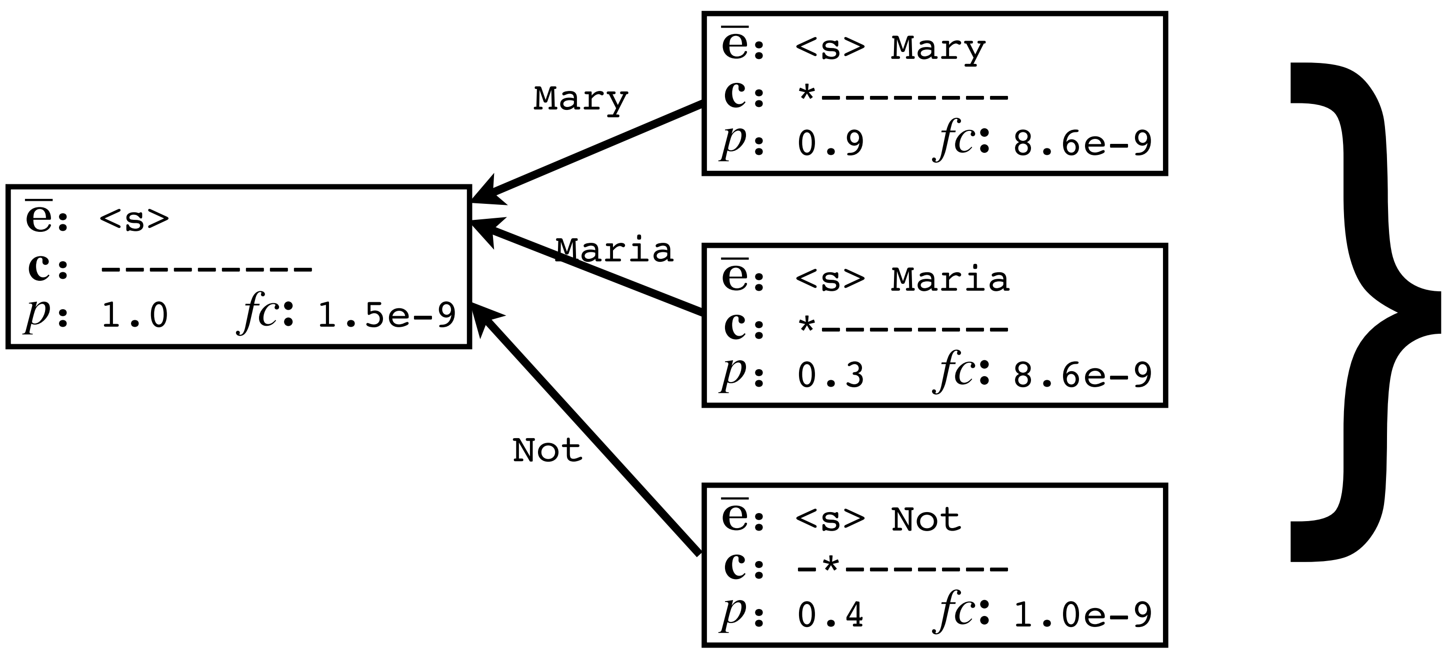
f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



f: Maria no dio una bofetada a la bruja verde

Q[0] Q[1] Q[2] ...



Future costs make these hypotheses comparable.

Decoding summary

- Finding the best hypothesis is NP-hard
  - Even with no language model, there are an exponential number of states!
  - Solution 1: limit reordering
  - Solution 2: (lossy) pruning