# Analyzing Errors of Unsupervised Learning

ACL 2008 Columbus, Ohio

June 18, 2008

Percy Liang Dan Klein

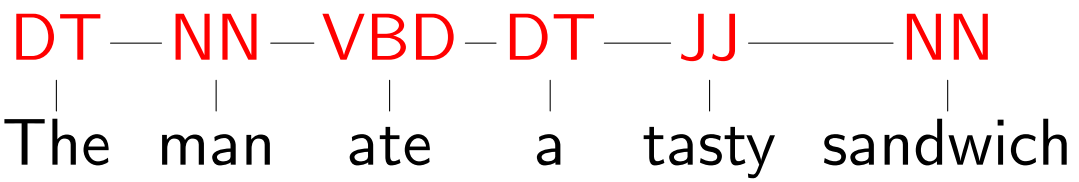
# Unsupervised grammar induction

Goal: induce hidden syntax

The man ate a tasty sandwich

# Unsupervised grammar induction

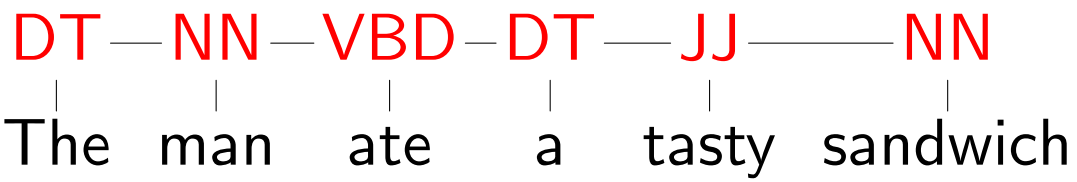
Goal: induce hidden syntax




POS tagging

# Unsupervised grammar induction

Goal: induce hidden syntax




DT NN VBD

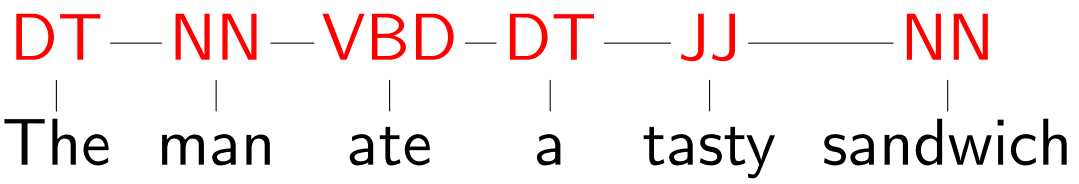
DT

JJ NN

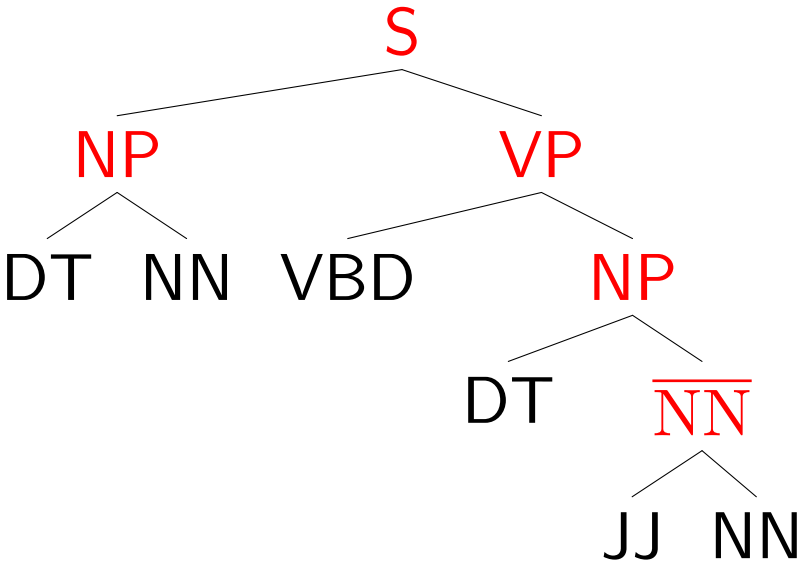
POS tagging

# Unsupervised grammar induction

Goal: induce hidden syntax



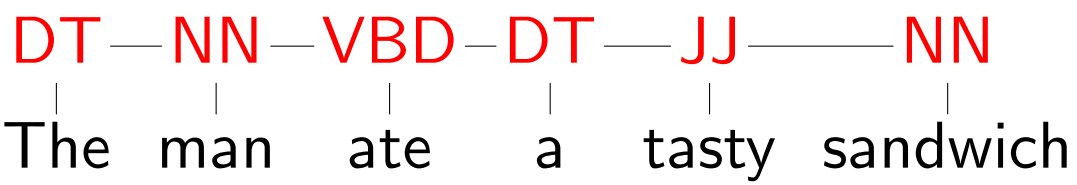

POS tagging



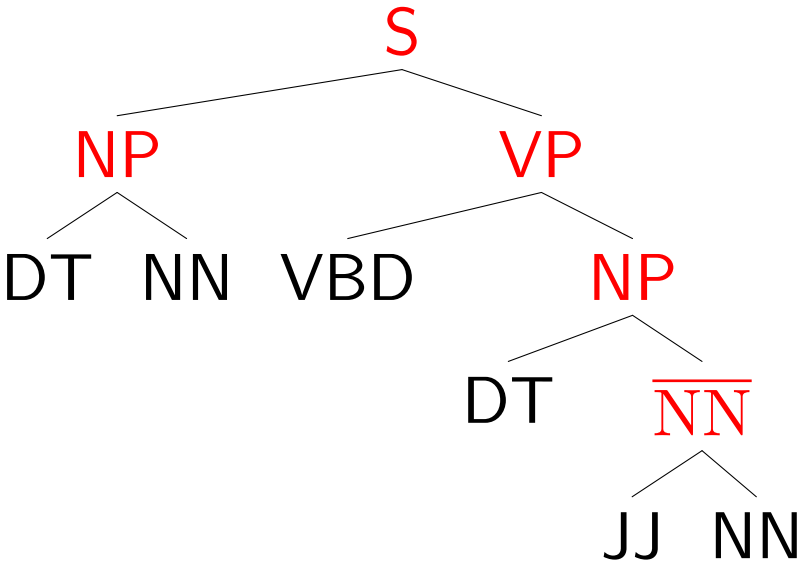

Constituency parsing

# Unsupervised grammar induction

Goal: induce hidden syntax




POS tagging



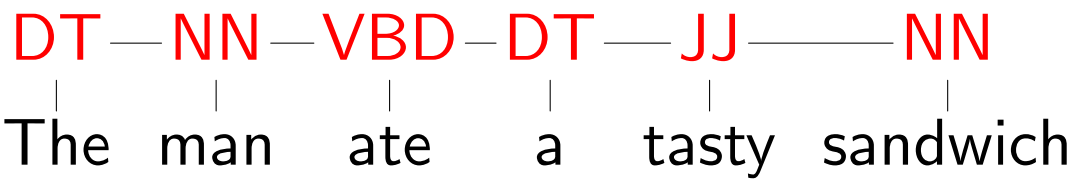

Constituency parsing

For example, on POS tagging using HMMs:

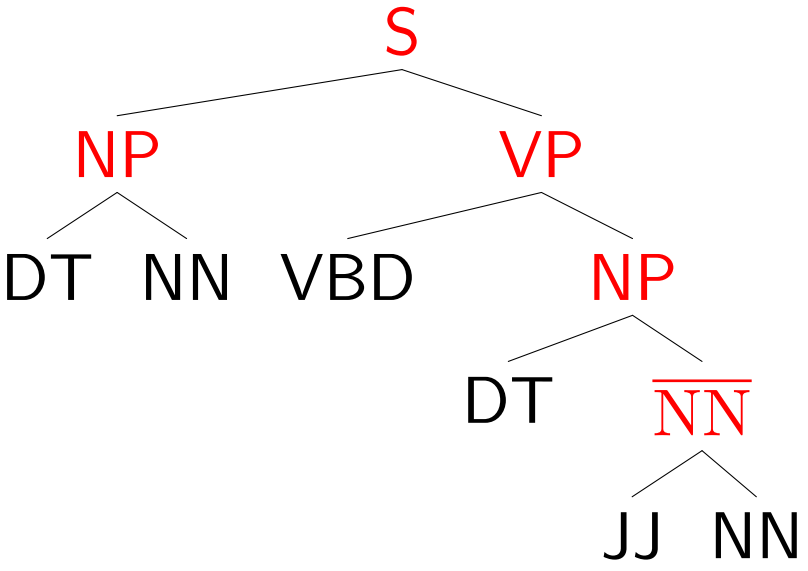
Unsupervised using EM $\approxeq 60\%$

# Unsupervised grammar induction

Goal: induce hidden syntax




POS tagging



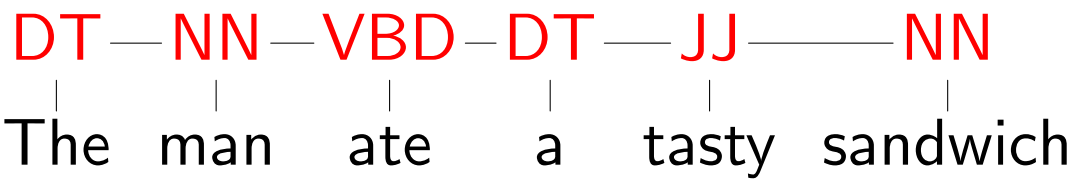

Constituency parsing

For example, on POS tagging using HMMs:

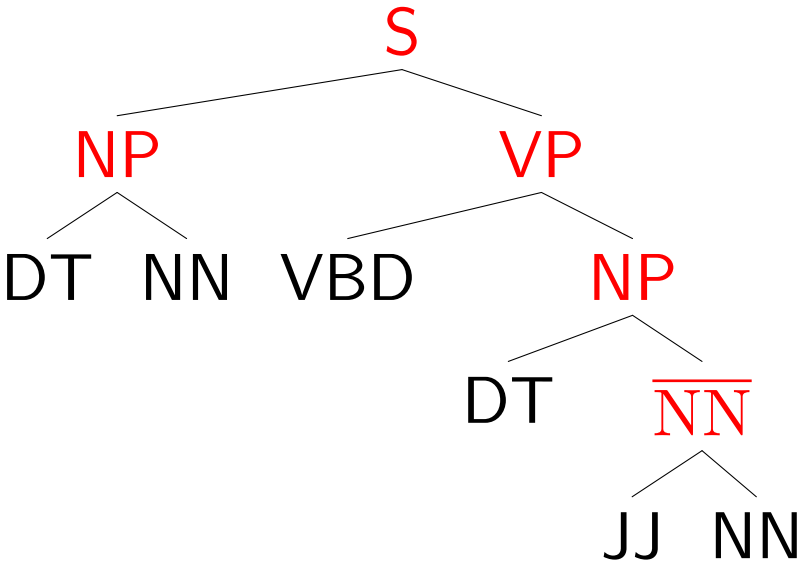
- Unsupervised using EM $\approxeq 60\%$
- Supervised $\geq 90\%$

# Unsupervised grammar induction

Goal: induce hidden syntax




POS tagging




Constituency parsing

For example, on POS tagging using HMMs:

- Unsupervised using EM $\approxeq 60\%$
- Supervised $\geq 90\%$

**Why does EM fail?**

# Four types of errors:

# Four types of errors:

- Optimization error
  - Local optima

# Four types of errors:

Optimization error
- Local optima

Estimation error
- Limited data

# Four types of errors:

Optimization error
- Local optima

Estimation error
- Limited data

Approximation error
- Likelihood objective $\not\Leftrightarrow$ accuracy

# Four types of errors:

## Optimization error

Local optima

## Estimation error

Limited data

## Approximation error

Likelihood objective $\not\Leftrightarrow$ accuracy

## Identifiability error

Different parameter settings $\longrightarrow$ same objective

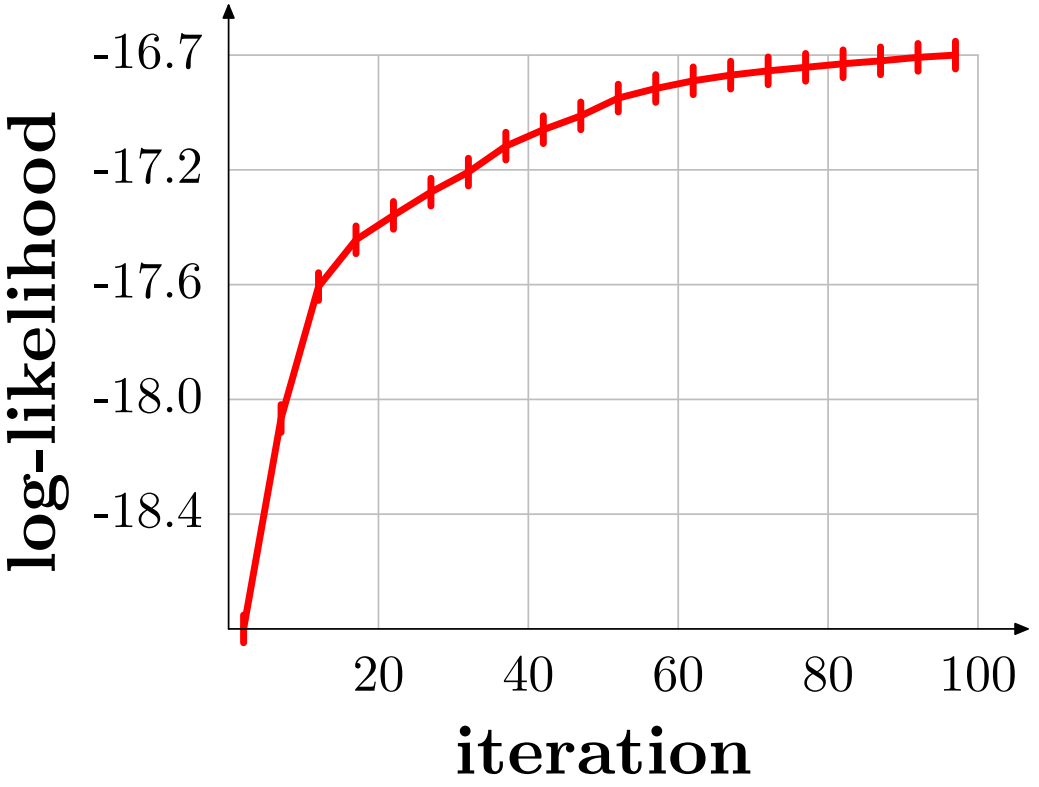
# Approximation error

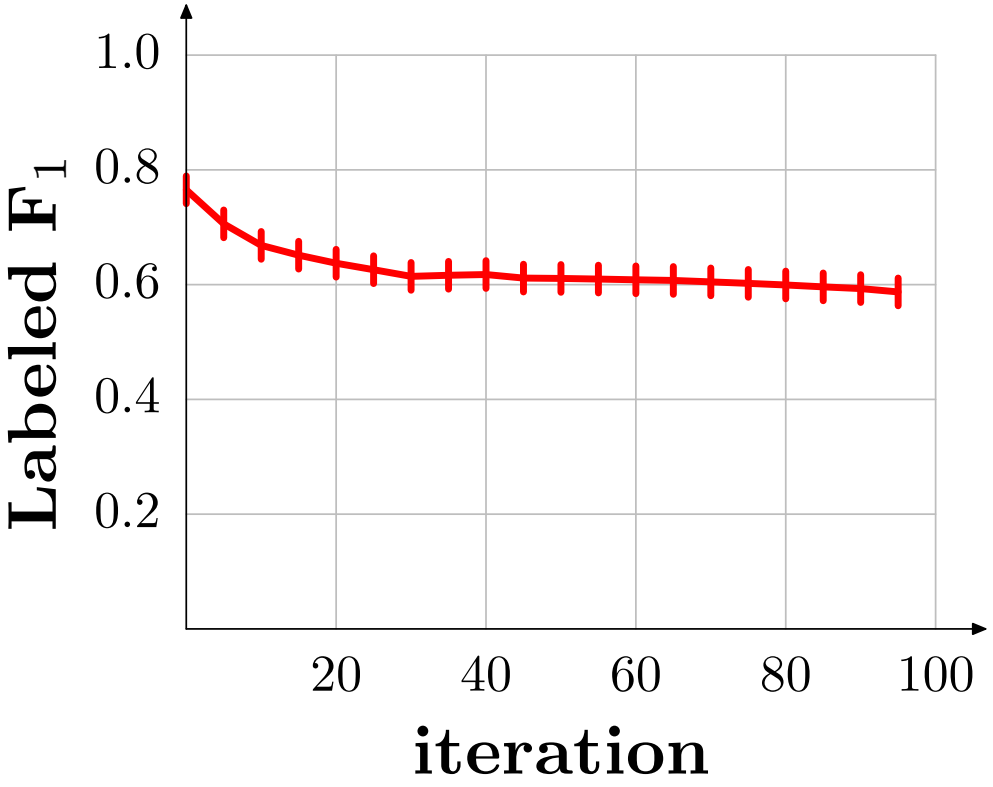
Problem: model likelihood $\not\Leftrightarrow$ prediction accuracy

# Approximation error

Problem: model likelihood $\not\Leftrightarrow$ prediction accuracy

PCFG (EM starting from supervised parameter estimate):

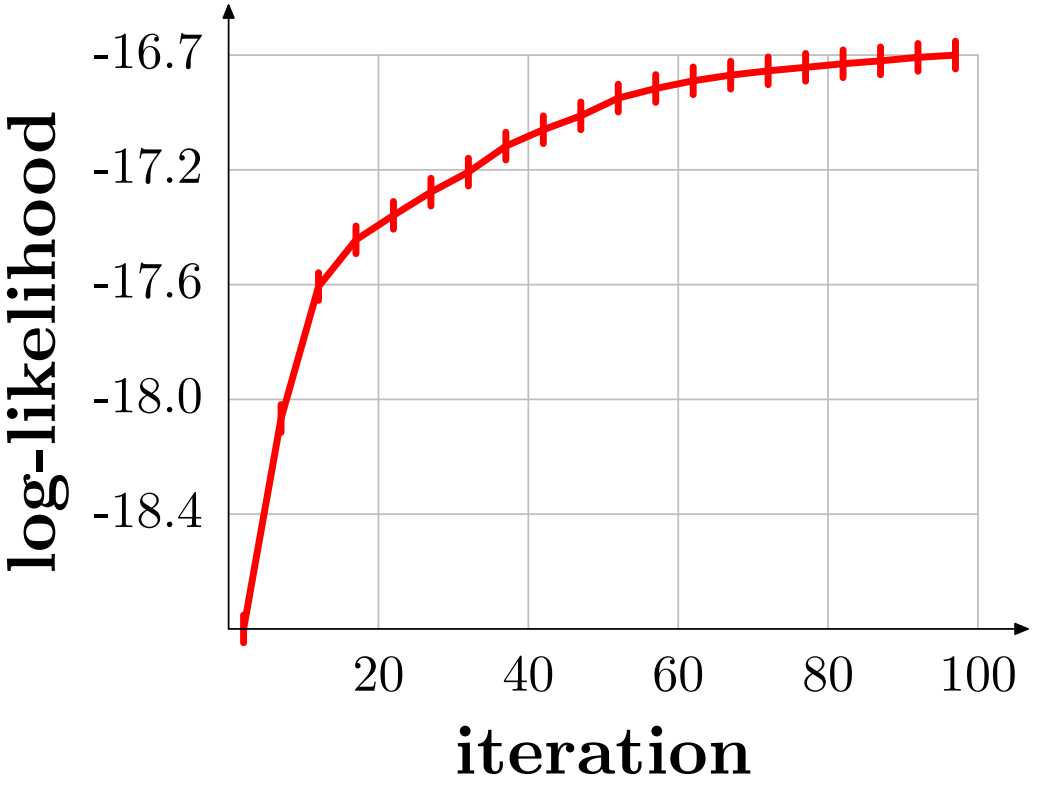


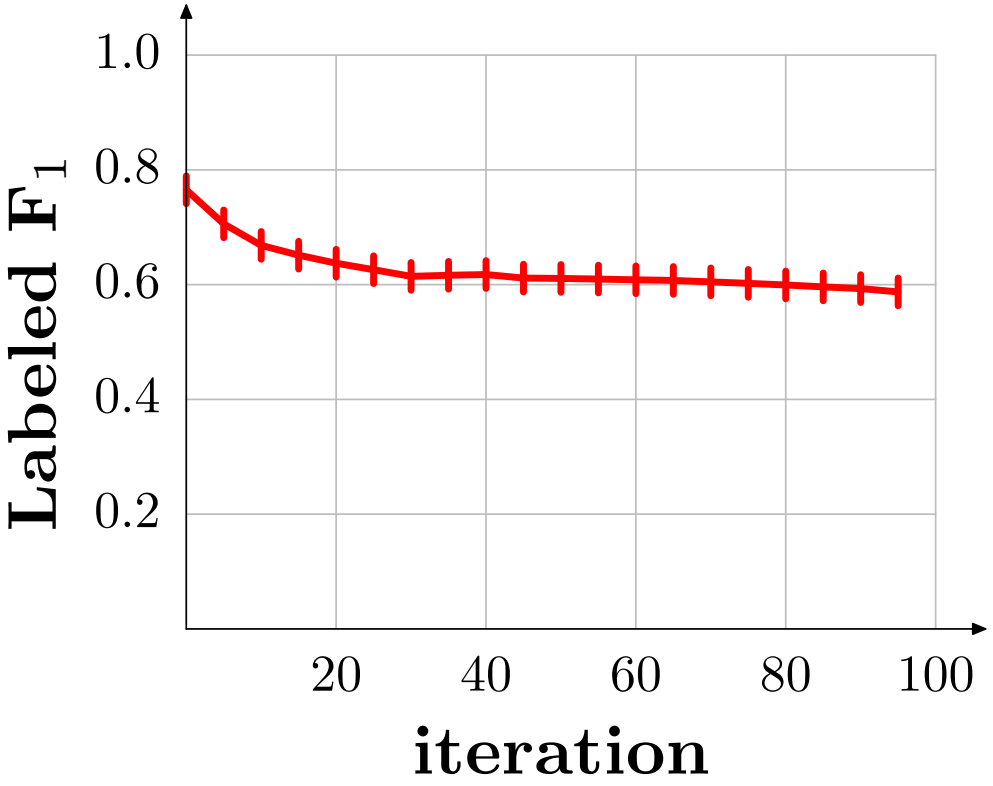



# Approximation error

Problem: model likelihood $\not\Leftrightarrow$ prediction accuracy

PCFG (EM starting from supervised parameter estimate):




**What qualitative changes is EM making?**

# Migrations

For the HMM:

Truth

DT NN NN RB VBD NNS
The chief executive allegedly made contributions

Iteration 1

DT JJ NN RB VBN NNS
The chief executive allegedly made contributions

# Migrations

For the HMM:

| | | | | | | |
|---|---|---|---|---|---|---|
| Truth | DT | NN | NN | RB | VBD | NNS |
| | The | chief | executive | allegedly | made | contributions |
| | | | | ↓ | | |
| Iteration 1 | DT | JJ | NN | RB | VBN | NNS |
| | The | chief | executive | allegedly | made | contributions |

Summarize changes by a set of migrations:

NN → JJ

VBD → VBN

# Migrations

For the HMM:

Truth

DT NN NN RB VBD NNS
The chief executive allegedly made contributions

Iteration 1

DT JJ NN RB VBN NNS
The chief executive allegedly made contributions

Summarize changes by a set of migrations:

NN→NN

JJ→NN

VBD→made

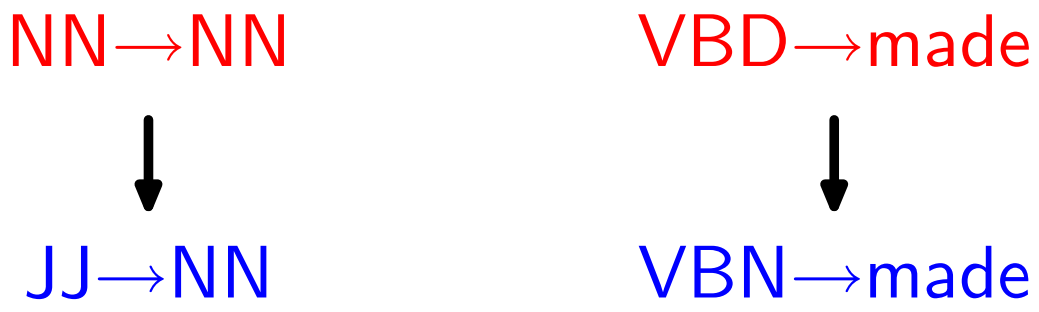
VBN→made

# Migrations

For the HMM:

| | | | | | | |
|---|---|---|---|---|---|---|
| Truth | DT | NN | NN | RB | VBD | NNS |
| | The | chief | executive | allegedly | made | contributions |
| | | | | ↓ | | |
| Iteration 1 | DT | JJ | NN | RB | VBN | NNS |
| | The | chief | executive | allegedly | made | contributions |

Summarize changes by a set of migrations:




**What are the prominent migrations over the entire corpora?**

# Top HMM migrations

**Iteration 1**

START → ~~NN~~ **NNP**

Sentence-initial nouns are often proper

$START\ Revenue/NN/NNP\ rose$

# Top HMM migrations

**Iteration 1**

START → ~~NN~~ **NNP**

Sentence-initial nouns are often proper

*START Revenue/NN/NNP rose*

~~NN~~ **JJ** → NN

Noun adjuncts → adjectives (inconsistent gold tags)

*chief/NN/JJ executive/NN officer*

# Top HMM migrations

**Iteration 1**

START → ~~NN~~ **NNP**

Sentence-initial nouns are often proper
*START Revenue/NN/NNP rose*

~~NN~~ **JJ** → NN

Noun adjuncts → adjectives (inconsistent gold tags)
*chief/NN/JJ executive/NN officer*

NNP → ~~NNP~~ **NNPS**

Inconsistent gold tags
*UBS Securities/NNP/NNPS*

# Top HMM migrations

**Iteration 1**

START → ~~NN~~ **NNP**

Sentence-initial nouns are often proper
*START Revenue/NN/NNP rose*

~~NN~~ **JJ** → NN

Noun adjuncts → adjectives (inconsistent gold tags)
*chief/NN/JJ executive/NN officer*

NNP → ~~NNP~~ **NNPS**

Inconsistent gold tags
*UBS Securities/NNP/NNPS*

**Iteration 2**

~~NN~~ **JJ** → NN

(same as above)

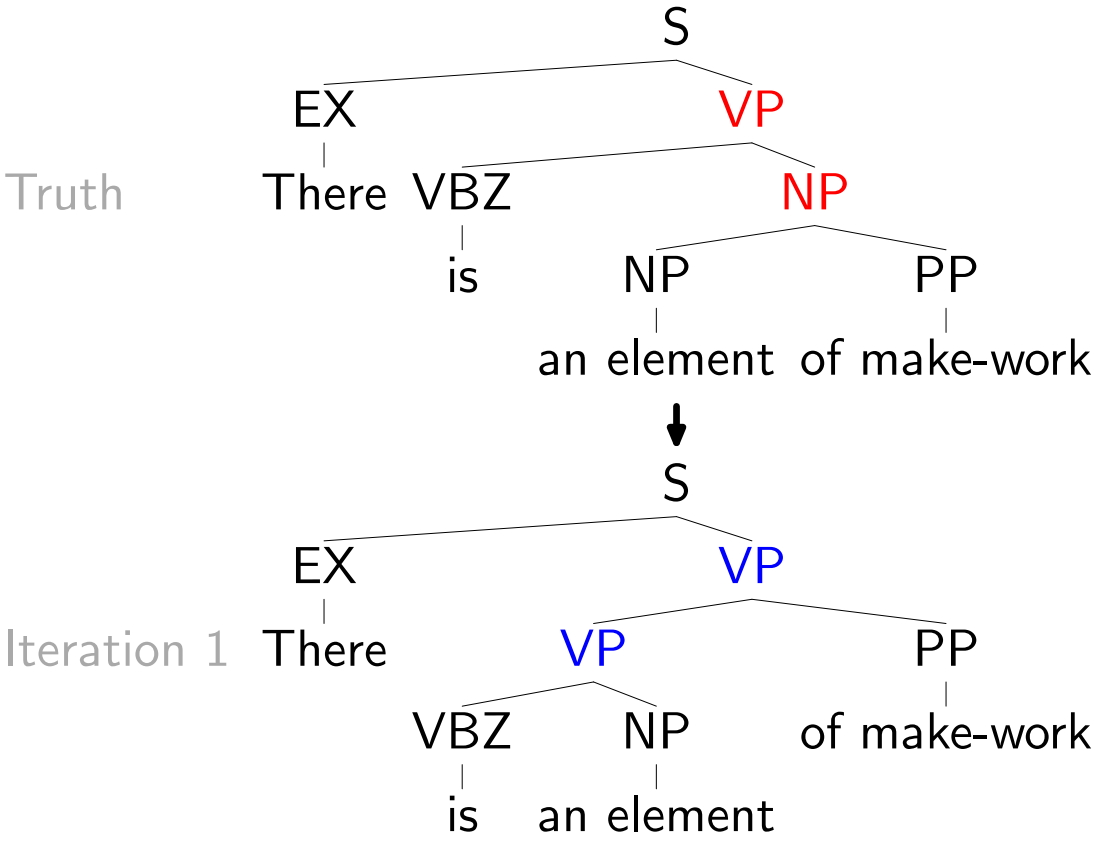
START → ~~NN~~ **NNP**

(same as above)

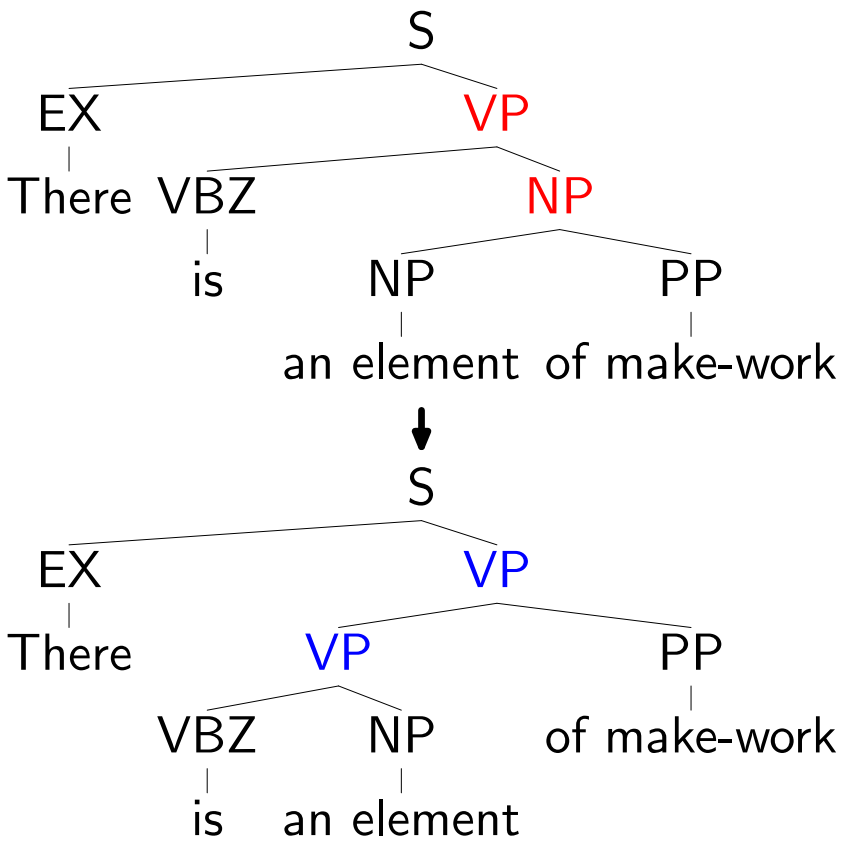
~~JJ~~ **RB** → TO

Inconsistent gold tags
*contribute much/JJ/RB to*

# Meta-modeling for PCFGs



# Meta-modeling for PCFGs




Truth

NP→NP PP

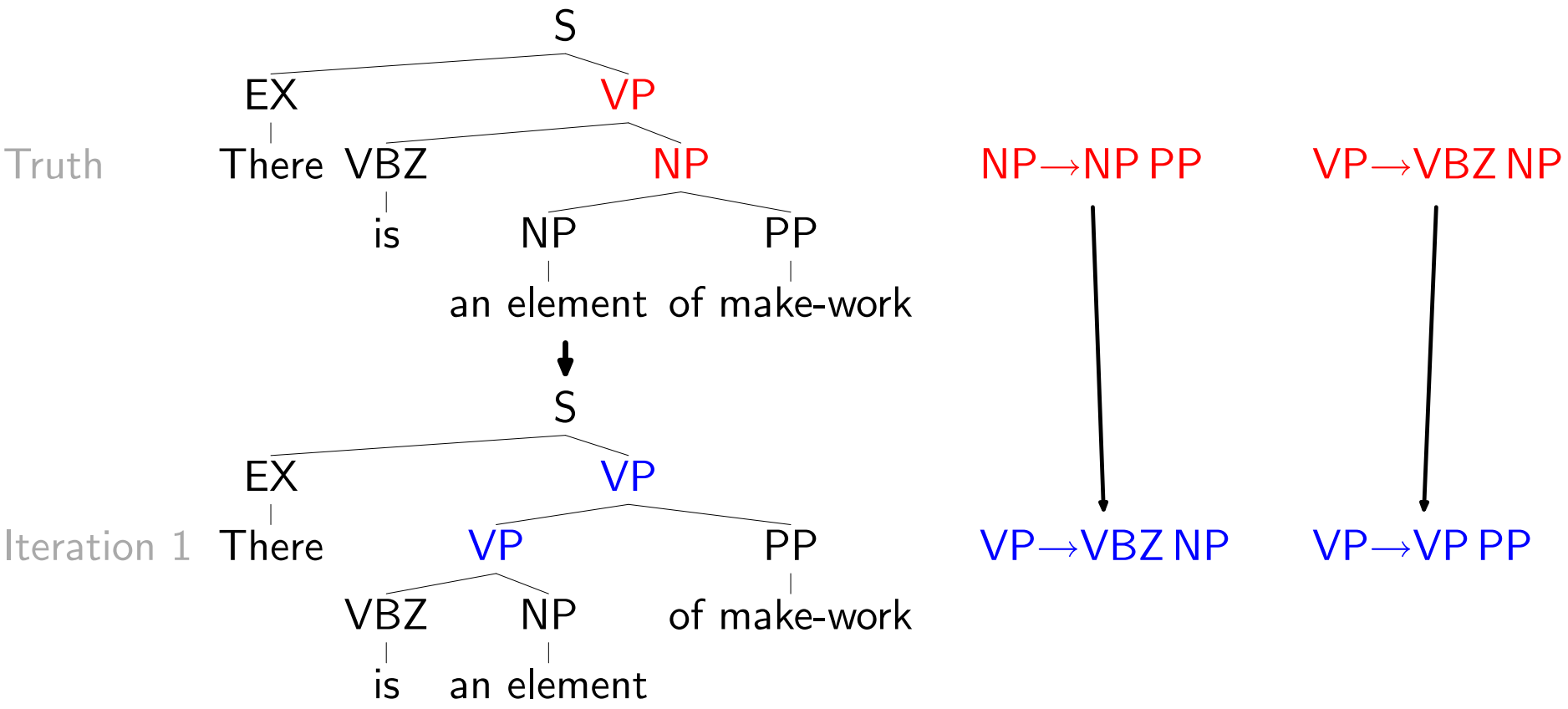
VP→VBZ NP

Iteration 1

VP→VBZ NP

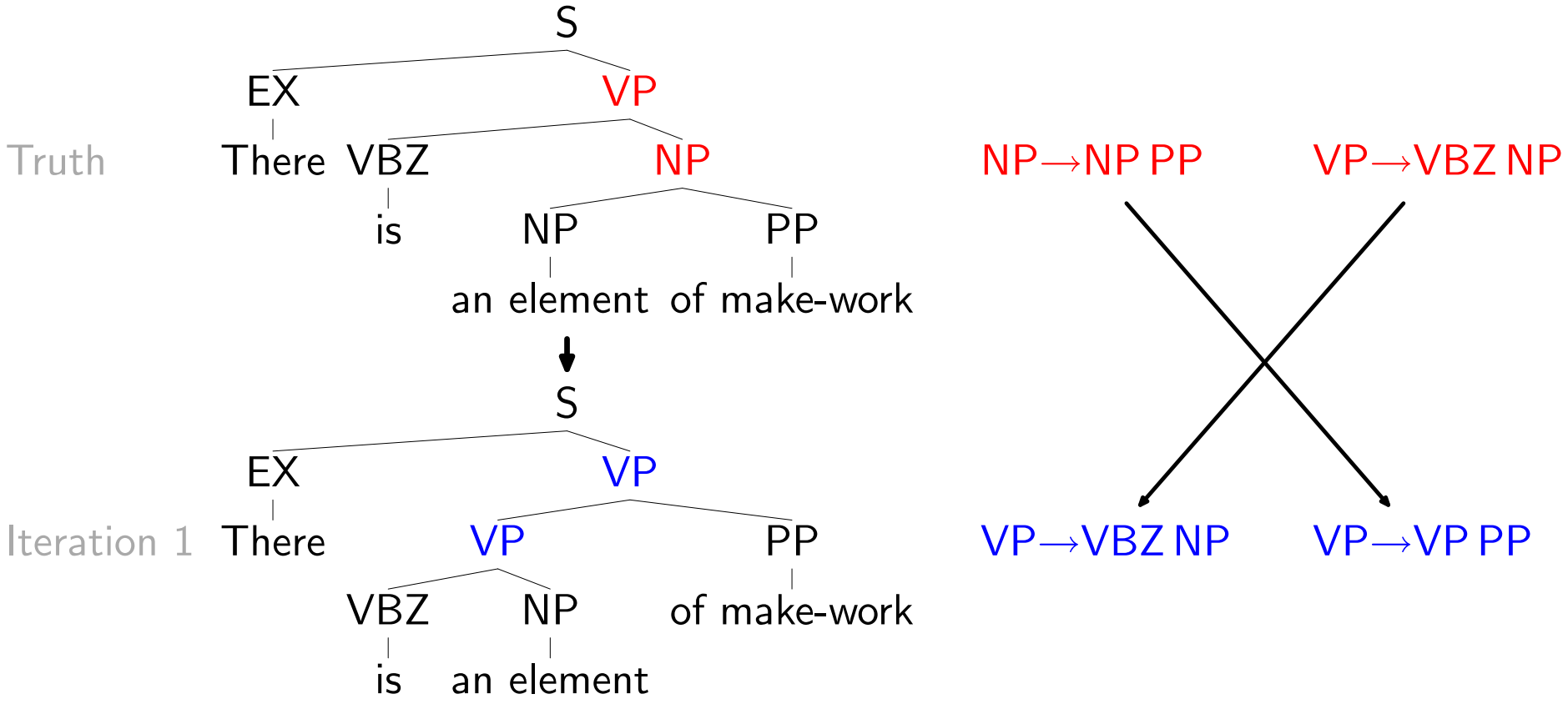
VP→VP PP

# Meta-modeling for PCFGs



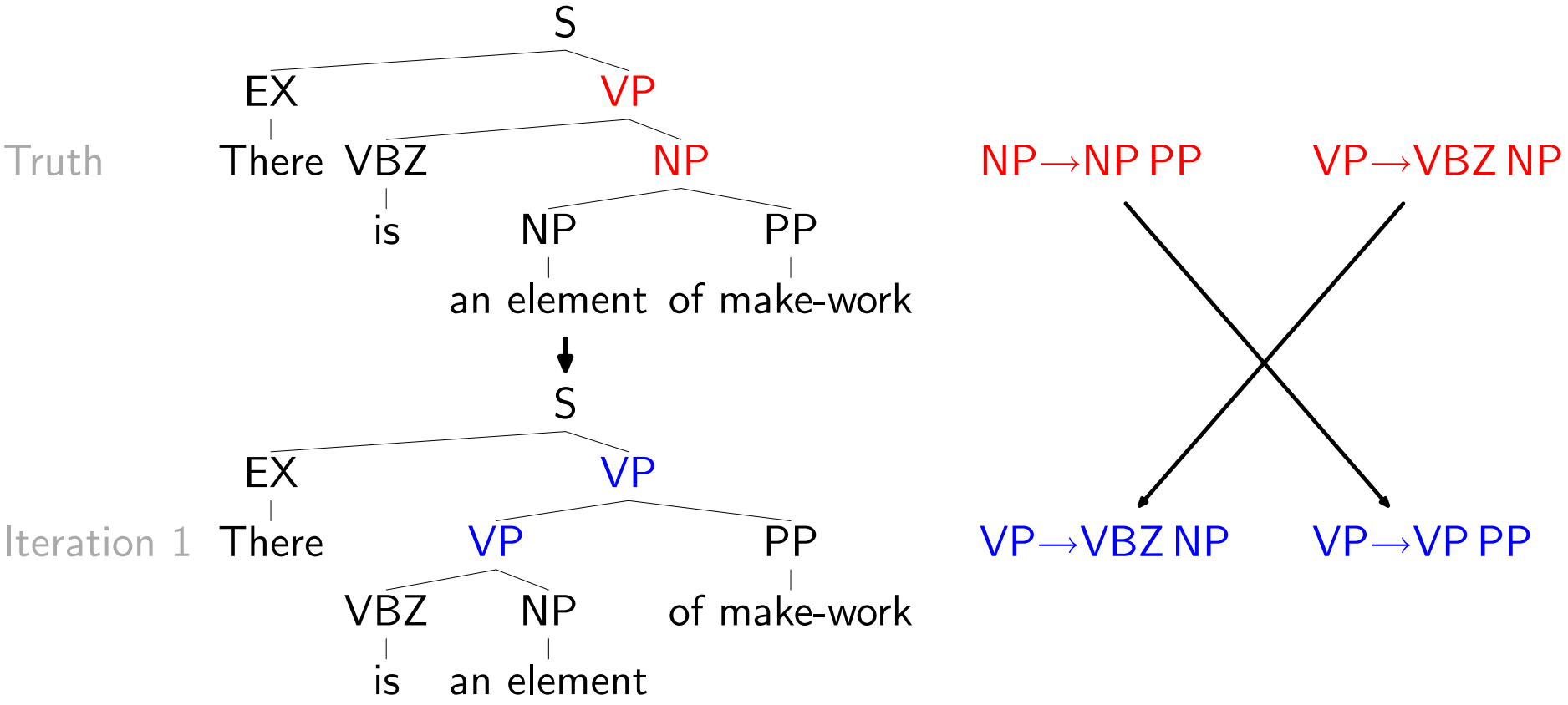

Migrations less clear due to uncertainty in tree structure...

# Meta-modeling for PCFGs




Migrations less clear due to uncertainty in tree structure...

# Meta-modeling for PCFGs

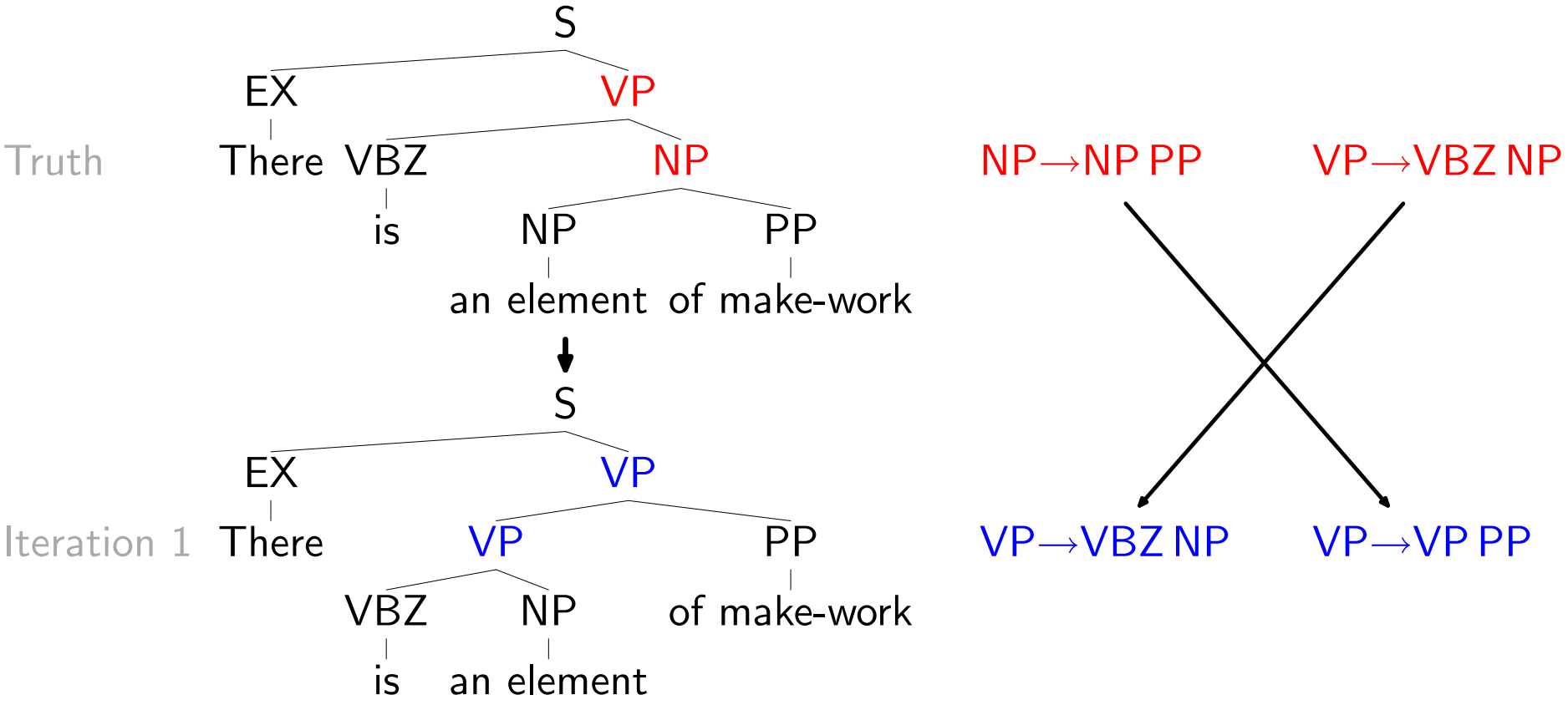



Migrations less clear due to uncertainty in tree structure...

## Our approach: use a meta-model

- Migrations are hidden alignments to be learned
- Fit using EM

# Meta-modeling for PCFGs



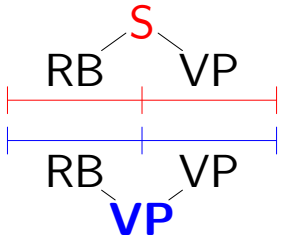

Migrations less clear due to uncertainty in tree structure...

## Our approach: use a meta-model

- Migrations are hidden alignments to be learned
- Fit using EM (convex, similar to IBM model 1)

# Top PCFG migrations learned by meta-model

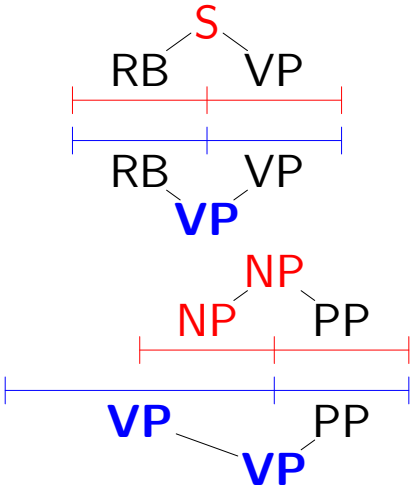
**Iteration 1**




Sentential adverbs → VP adverbs

# Top PCFG migrations learned by meta-model

**Iteration 1**

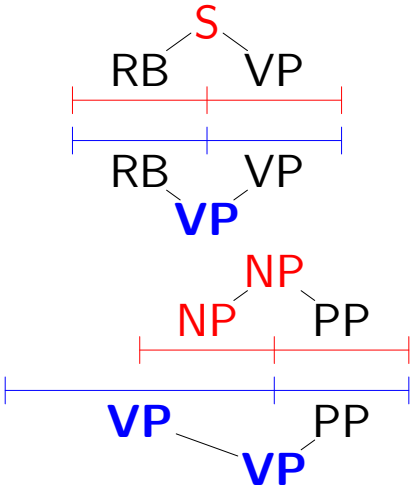



Sentential adverbs → VP adverbs

PPs raised from NPs to verbal level

# Top PCFG migrations learned by meta-model

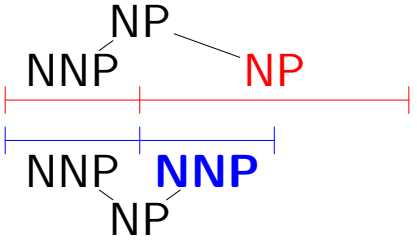
**Iteration 1**




Sentential adverbs → VP adverbs

PPs raised from NPs to verbal level

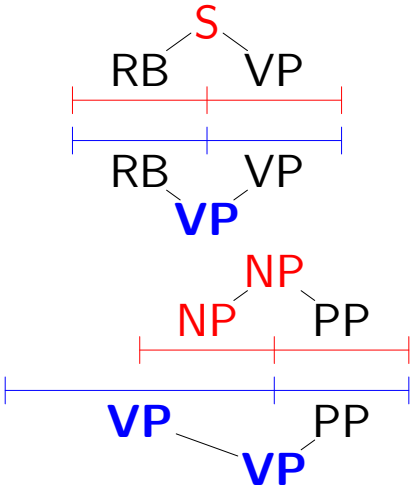
**Iteration 2**




Right-branching → left-branching structures

# Top PCFG migrations learned by meta-model

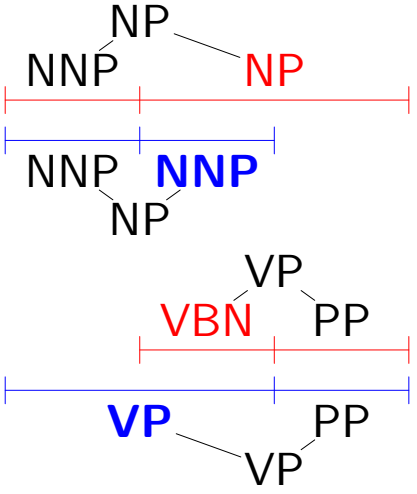
**Iteration 1**




Sentential adverbs → VP adverbs

PPs raised from NPs to verbal level

**Iteration 2**




Right-branching → left-branching structures

PP raised to higher VP

# Meta-modeling summary

- Meta-model: a diagnostic tool to analyze errors systematically

# Meta-modeling summary

- Meta-model: a diagnostic tool to analyze errors systematically
- General phenomenon: regularization of syntactic structure

# Meta-modeling summary

- Meta-model: a diagnostic tool to analyze errors systematically
- General phenomenon: regularization of syntactic structure

✓ Approximation error
Identifiability error
Estimation error
Optimization error

# Identifiability error

$\mathbf{x}$: input sentence

$\mathbf{y}$: hidden output

$p_\theta(\mathbf{x}, \mathbf{y})$: joint distribution with parameters $\theta$

# Identifiability error

$\mathbf{x}$: input sentence

$\mathbf{y}$: hidden output

$p_\theta(\mathbf{x}, \mathbf{y})$: joint distribution with parameters $\theta$

Non-identifiability:

$\boxed{\theta_1}$ ? $\boxed{\theta_2}$

# Identifiability error

$\mathbf{x}$: input sentence

$\mathbf{y}$: hidden output

$p_\theta(\mathbf{x}, \mathbf{y})$: joint distribution with parameters $\theta$

Non-identifiability:

Learning is indifferent...

$$p_{\theta_1}(\mathbf{x}) = p_{\theta_2}(\mathbf{x})$$

$$\boxed{\theta_1} \quad ? \quad \boxed{\theta_2}$$

# Identifiability error

$\mathbf{x}$: input sentence

$\mathbf{y}$: hidden output

$p_\theta(\mathbf{x}, \mathbf{y})$: joint distribution with parameters $\theta$

Non-identifiability:

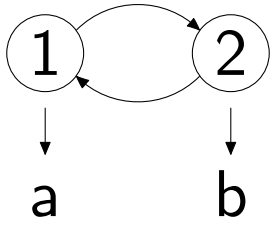
Learning is indifferent... $p_{\theta_1}(\mathbf{x}) = p_{\theta_2}(\mathbf{x})$

$\boxed{\theta_1} \;?\; \boxed{\theta_2}$

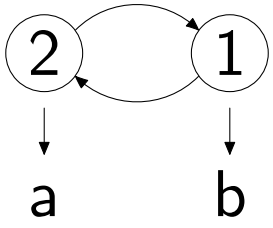
but matters to prediction (bad!) $p_{\theta_1}(\mathbf{y} \mid \mathbf{x}) \neq p_{\theta_2}(\mathbf{y} \mid \mathbf{x})$

# Examples of non-identifiability

- Label symmetries



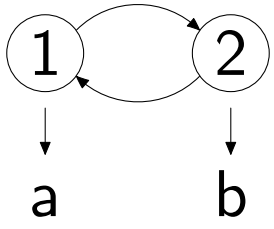

and



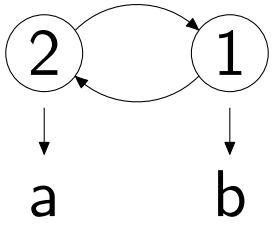

both generate abababab...

# Examples of non-identifiability

- Label symmetries

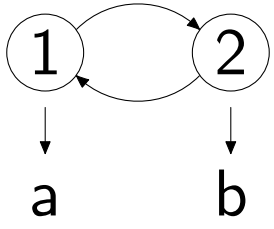



and

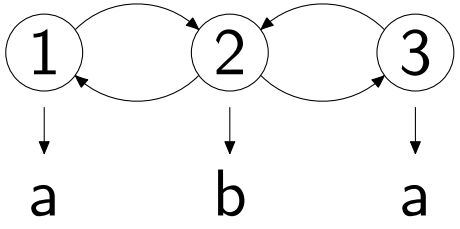



both generate abababab...

- $K$-state HMM (if true distribution is $< K$-state HMM)



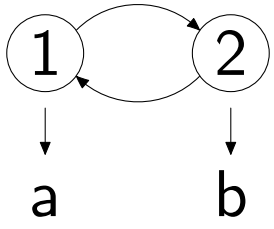

and



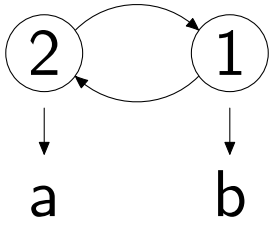

both generate abababab...

# Examples of non-identifiability

- Label symmetries

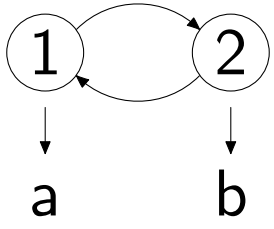



and

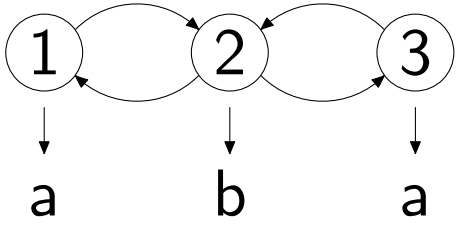



both generate abababab...

- $K$-state HMM (if true distribution is $< K$-state HMM)

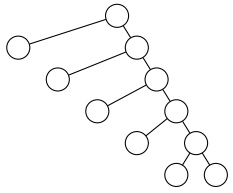



and

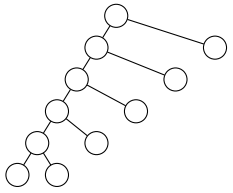



both generate abababab...

- PCFG (if true distribution is HMM)



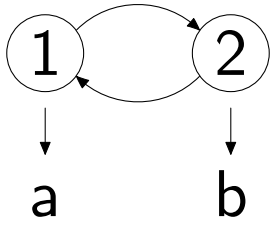
and

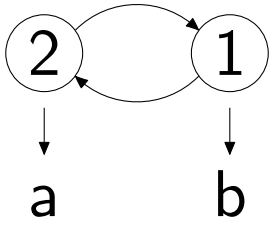


can both simulate the HMM

# Examples of non-identifiability

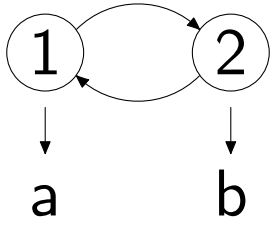
- Label symmetries

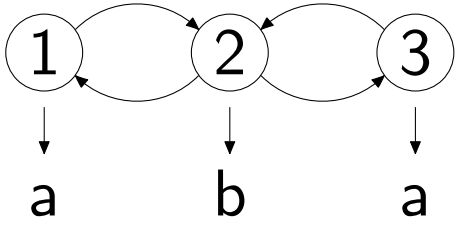


and


both generate abababab...

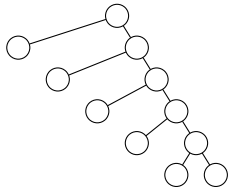
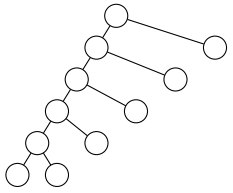
- $K$-state HMM (if true distribution is $< K$-state HMM)



and


both generate abababab...

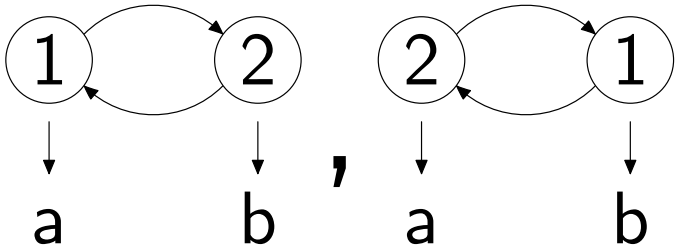
- PCFG (if true distribution is HMM)


and

can both simulate the HMM

Real data is complex, so last two are not an issue

# Identifiability and distance

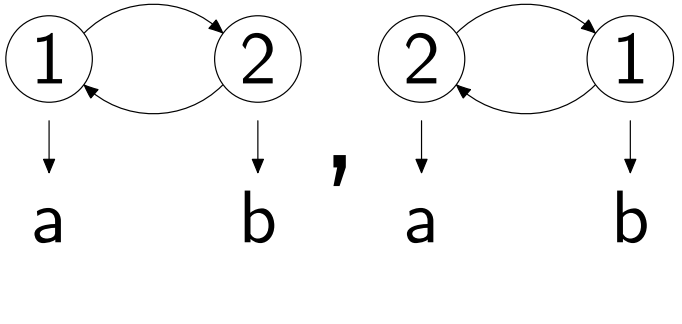
Given $\theta_1$ and $\theta_2$, how to measure distance between them?

Want distance $\left( \right.$ 

$\left. \right) = 0$

# Identifiability and distance

Given $\theta_1$ and $\theta_2$, how to measure distance between them?

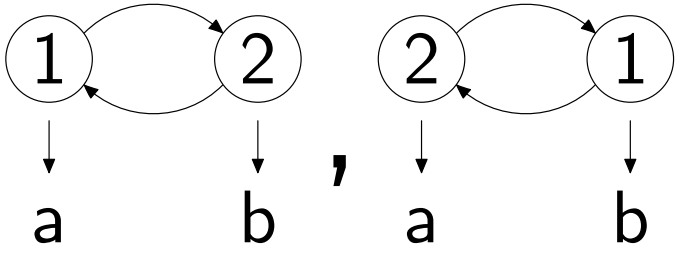
Want distance $\left( \right.$ 

$\left. \right) = 0$

- Computing label-permutation invariant distance is NP-hard
- We use bipartite matching to find lower and upper bounds

# Identifiability and distance

Given $\theta_1$ and $\theta_2$, how to measure distance between them?

Want distance $\left( \right.$ 

$\left. \right) = 0$

- Computing label-permutation invariant distance is NP-hard
- We use bipartite matching to find lower and upper bounds

✓ Approximation error
✓ Identifiability error
Estimation error
Optimization error

# Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM?

# Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM?

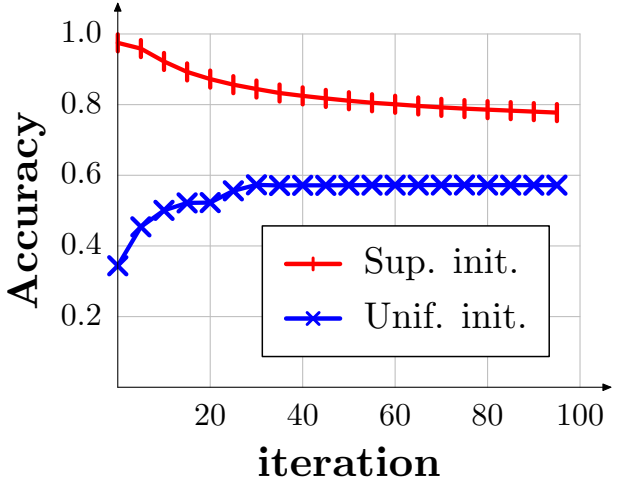
# Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM? No?

$\theta^*$ ← estimation error → global optimum ← optimization error → EM solution

HMM on 5K examples:

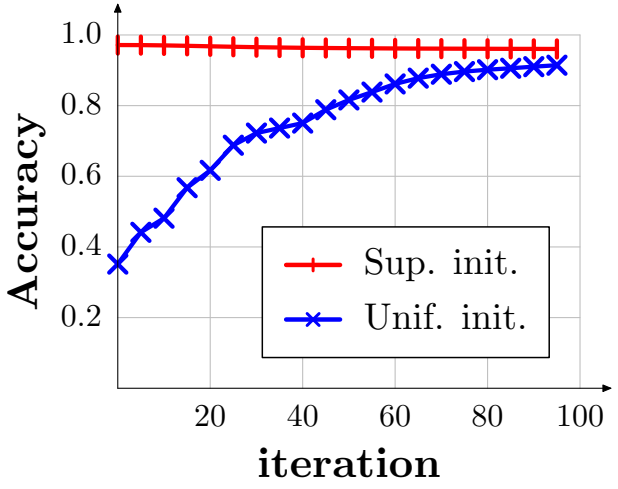


# Estimation and optimization errors

Experiment setup:

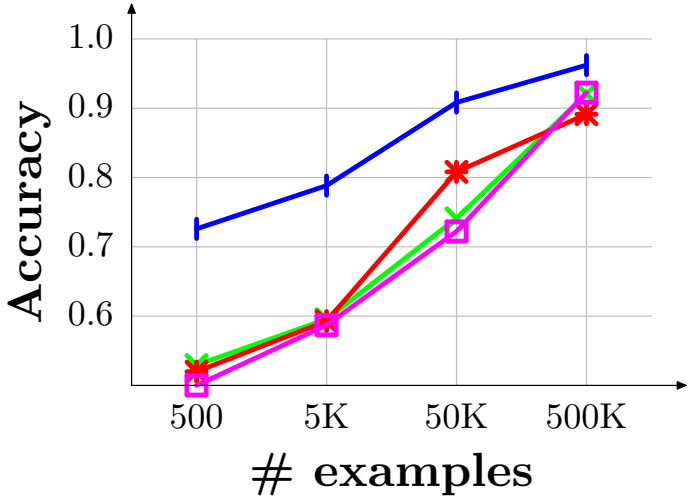
- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM? Yes!


HMM on 500K examples:



# Optimization error decreases with more data

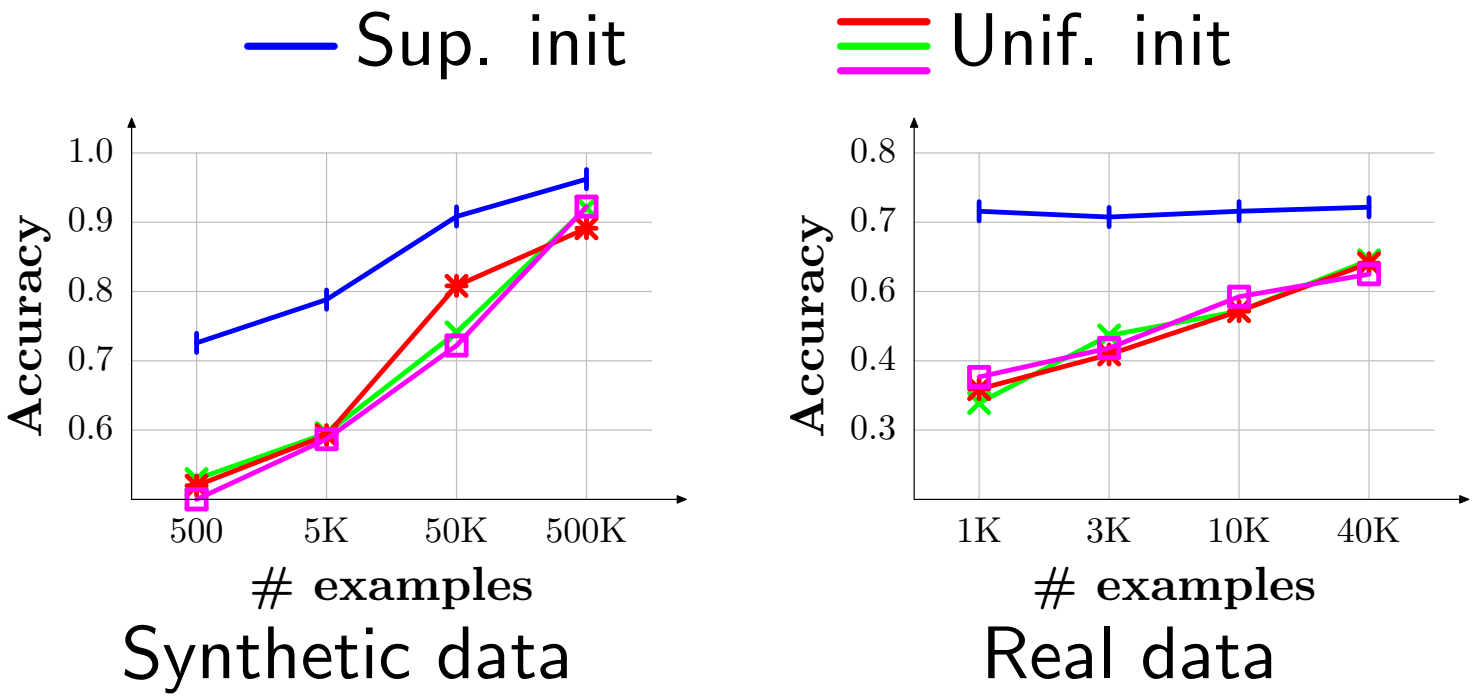
On HMM model (similar for PCFG and a dependency model):




Synthetic data

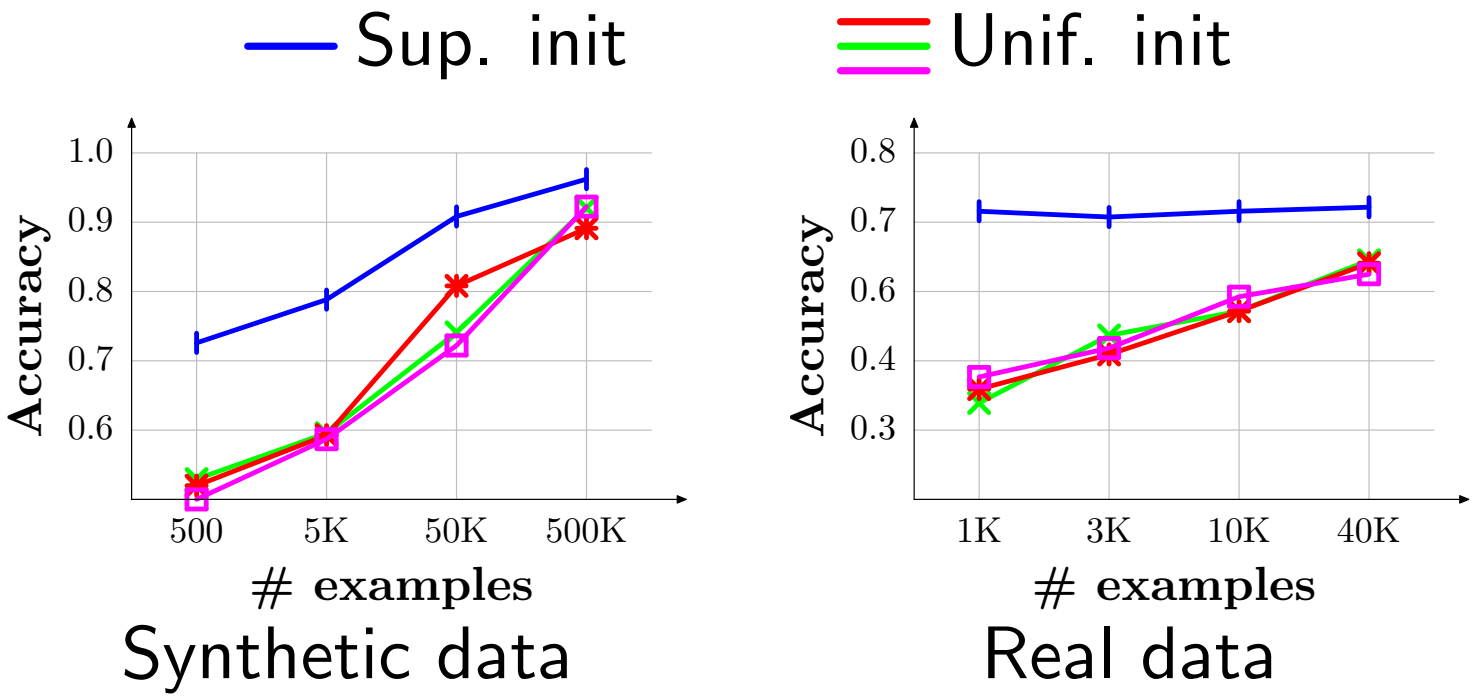
# Optimization error decreases with more data

On HMM model (similar for PCFG and a dependency model):



# Optimization error decreases with more data

On HMM model (similar for PCFG and a dependency model):




Why does this phenomenon happen?

- Intuition: with more data, EM can pick up the salient patterns more easily
- Was also shown for mixture of Gaussians [Srebro, 2006]

# Summary

- ✓ Approximation error
  - Meta-model: tool for systematic error analysis

# Summary

- ✓ Approximation error
  - Meta-model: tool for systematic error analysis
- ✓ Identifiability error
  - Distance robust to label symmetries

# Summary

- ✓ Approximation error
  - Meta-model: tool for systematic error analysis
- ✓ Identifiability error
  - Distance robust to label symmetries
- ✓ Estimation error
  - Decreases with more data

# Summary

- ✓ Approximation error
  - Meta-model: tool for systematic error analysis
- ✓ Identifiability error
  - Distance robust to label symmetries
- ✓ Estimation error
  - Decreases with more data
- ✓ Optimization error
  - Decreases with more data!