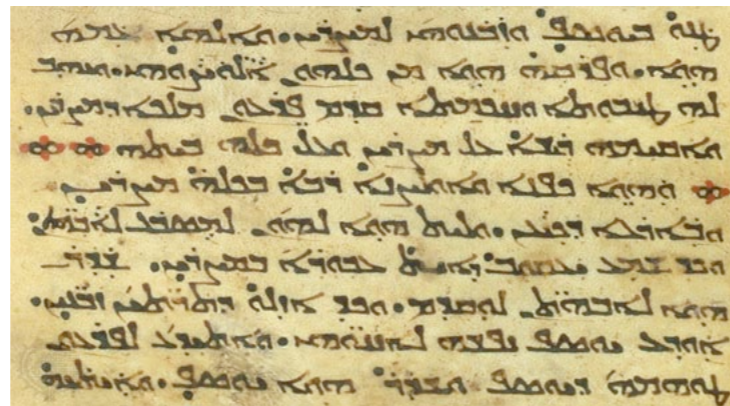
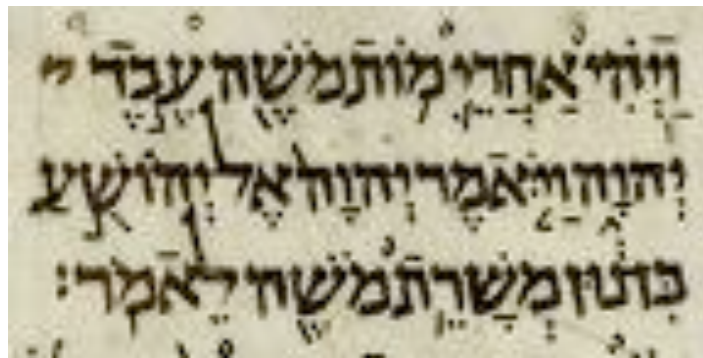


# Unsupervised Multilingual Learning for Morphological Segmentation



Benjamin Snyder and Regina Barzilay  
MIT

# Breaking the Unsupervised Ceiling

# Breaking the Unsupervised Ceiling

- Unsupervised models for many core tasks proposed

# Breaking the Unsupervised Ceiling

- Unsupervised models for many core tasks proposed
- Performance lags behind supervised models

# Breaking the Unsupervised Ceiling

- Unsupervised models for many core tasks proposed
- Performance lags behind supervised models

How can we cut the unsupervised/supervised performance gap?

# Multilingual Learning

## Goal:

Induce individual language structures  
*along* with interlingual connections with  
no supervision

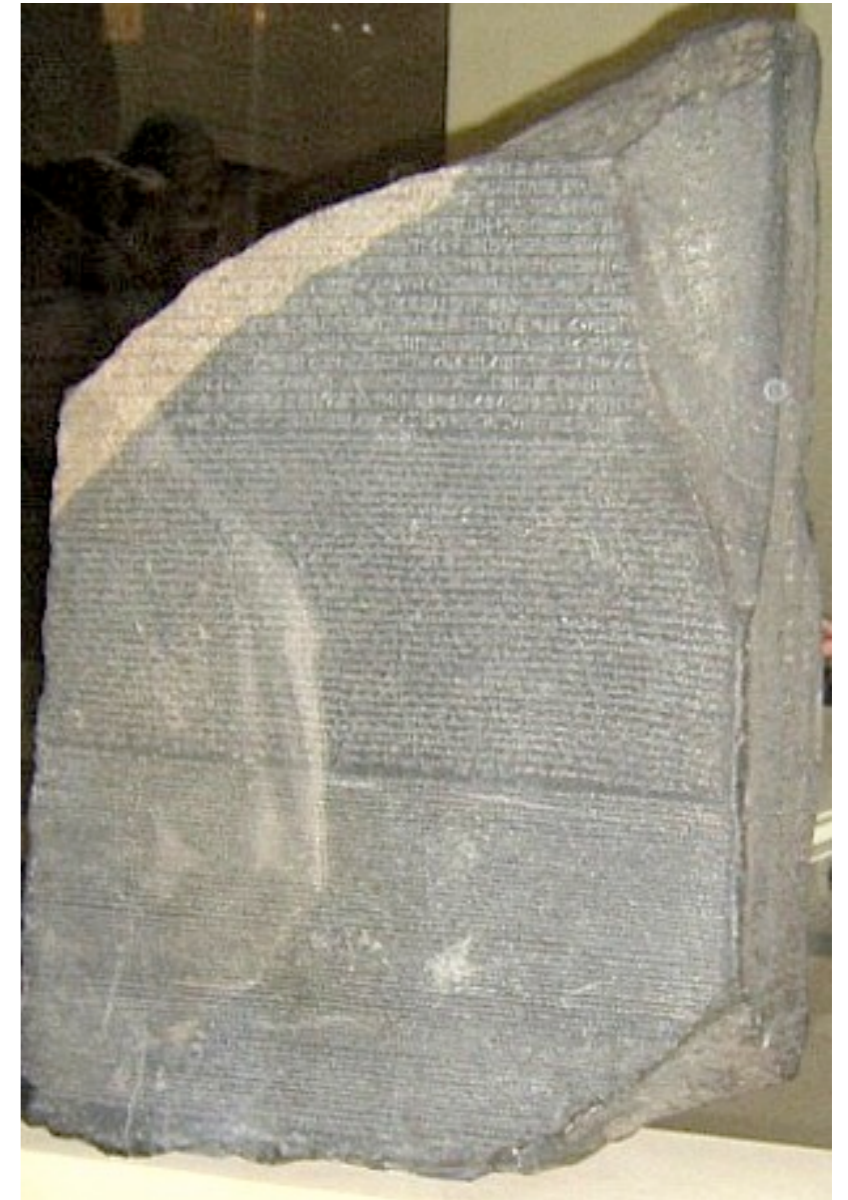
## Linguistic Motivation:

- Languages *related* structurally and historically
- But *differ systematically* in patterns of ambiguity and expression

# Historical Multilingual Learning

- Comparative study of languages goes back to Antiquity
- Achievements:
  - ▶ Deciphering unknown scripts
  - ▶ Understanding dead languages
  - ▶ Reconstructing proto-languages

Use language *similarity* as bridge  
Learn from language *difference*



Rosetta Stone

# Multilingual Corpora: A Rosetta Stone for Unsupervised NLP

בראשית ברא אלהים את השמים ואת הארץ

בֹּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والارض

“In the beginning God created the heavens and the earth...”



# Multilingual Corpora: A Rosetta Stone for Unsupervised NLP

בראשית ברא אלהים את השמים ואת הארץ  
בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ  
في البدء خلق الله السموات والارض

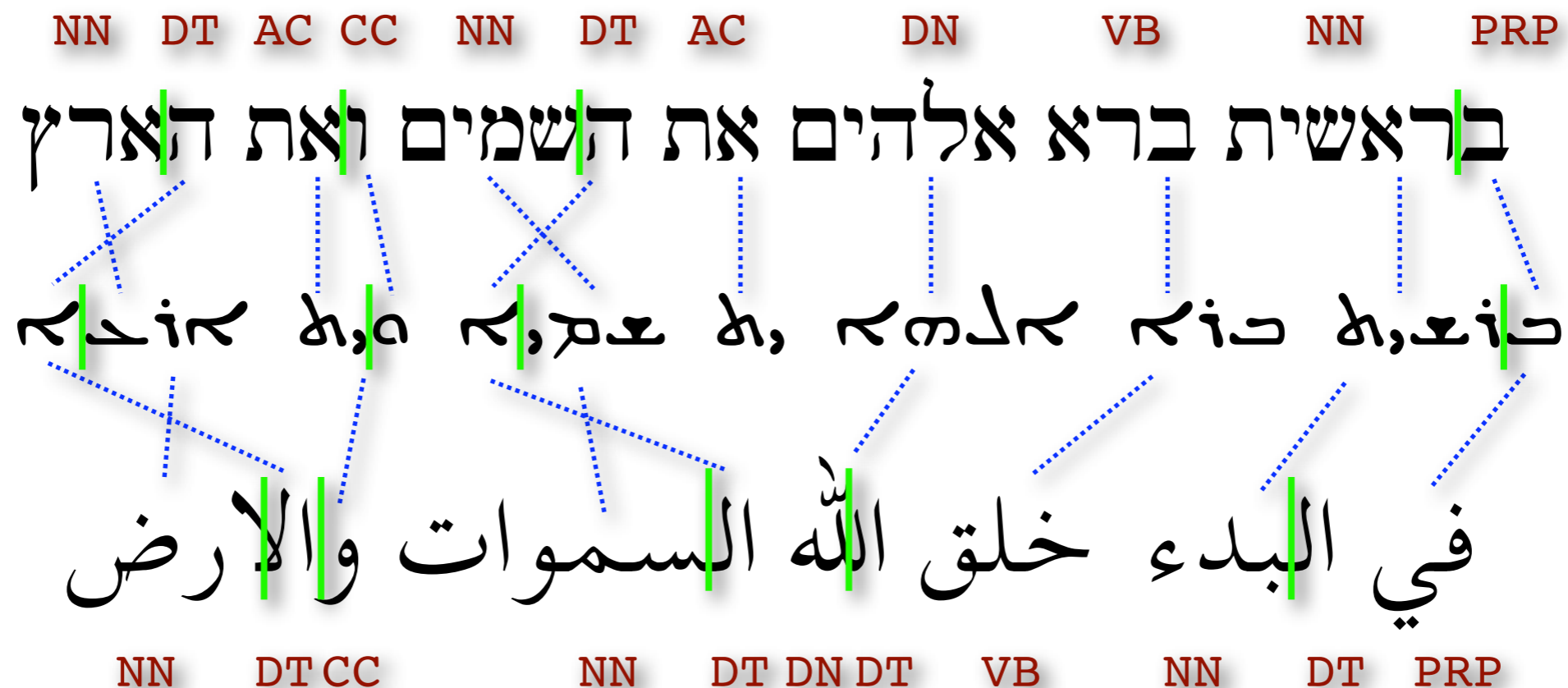
“In the beginning God created the heavens and the earth...”

# Multilingual Corpora: A Rosetta Stone for Unsupervised NLP

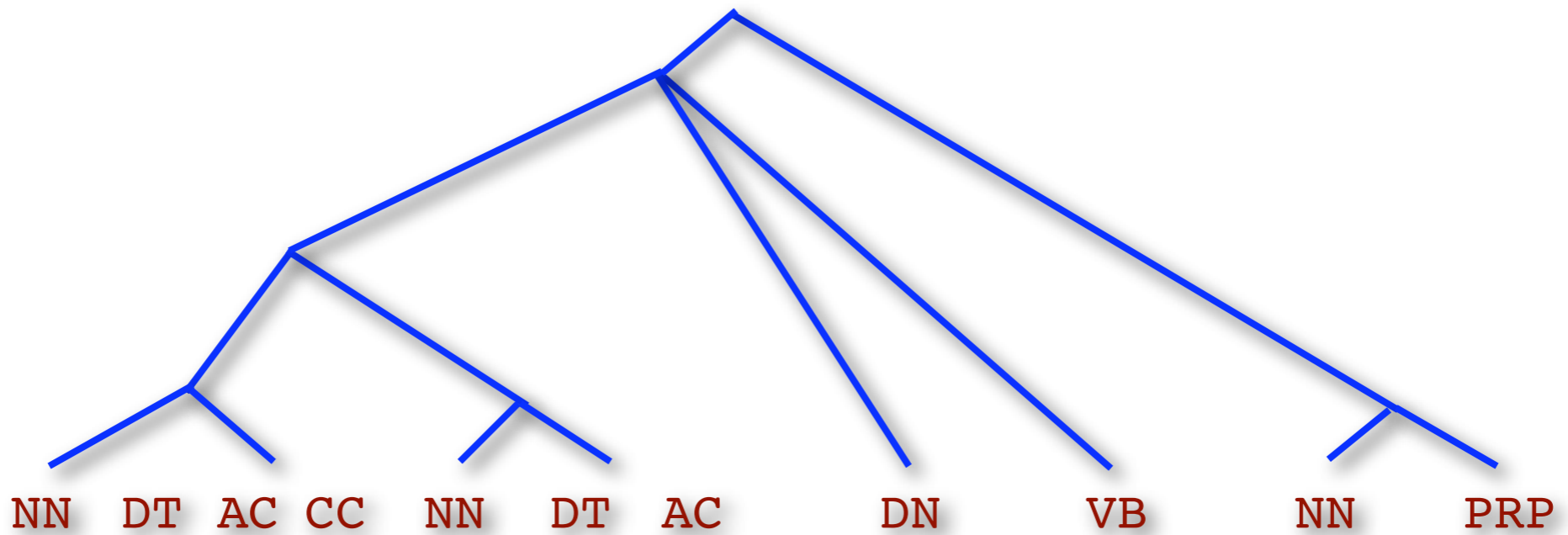
בראשית ברא אלהים את השמים ואת הארץ  
בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ  
בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ  
في البدء خلق الله السموات والارض

“In the beginning God created the heavens  
and the earth...”

# Multilingual Corpora: A Rosetta Stone for Unsupervised NLP



“In the beginning God created the heavens and the earth...”



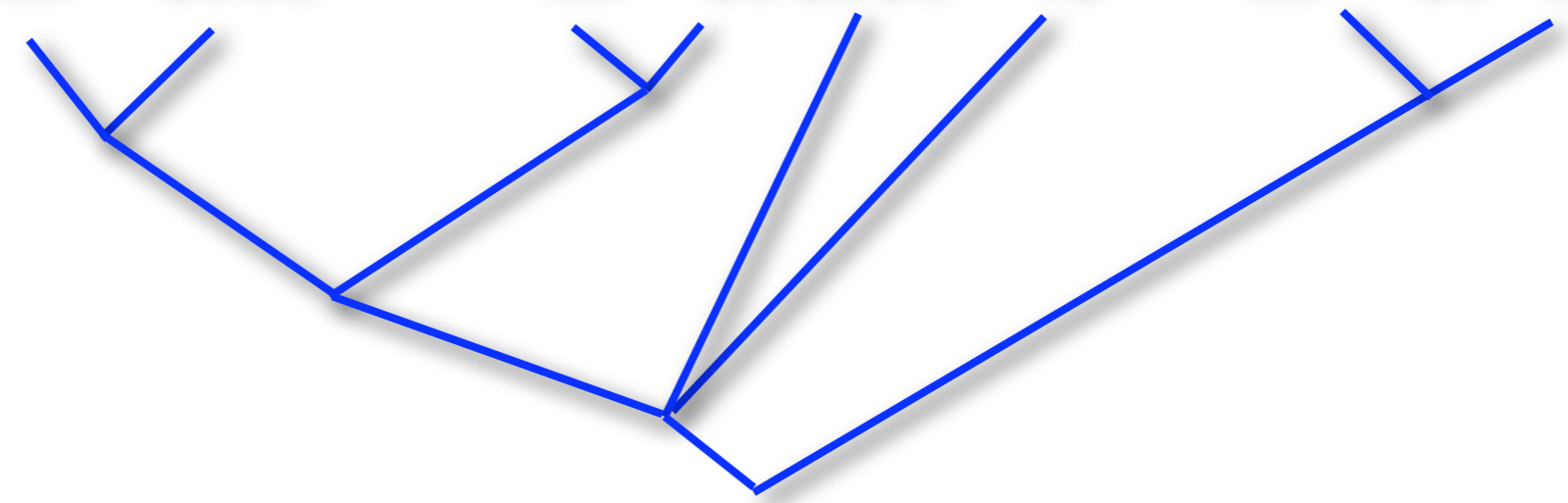
NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP



בראשית ברא אלהים את השמים ואת הארץ

בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والأرض

# Task: Morphological Segmentation

Input: raw bilingual parallel corpus with no additional resources or knowledge of the languages

Goal: segment each word into smallest units of meaning -- morphemes. e.g.,

fakatabuuhu

so they wrote it

# Task: Morphological Segmentation

Input: raw bilingual parallel corpus with no additional resources or knowledge of the languages

Goal: segment each word into smallest units of meaning -- morphemes. e.g.,

fakatabuuhu → fa|katab|uu|hu  
so they wrote it → so they wrote it

barş

byom

arşnu



in a land

barṣ

in a day

byom

our land

arṣnu

- Learn from different ambiguity patterns

in a land

b|arş

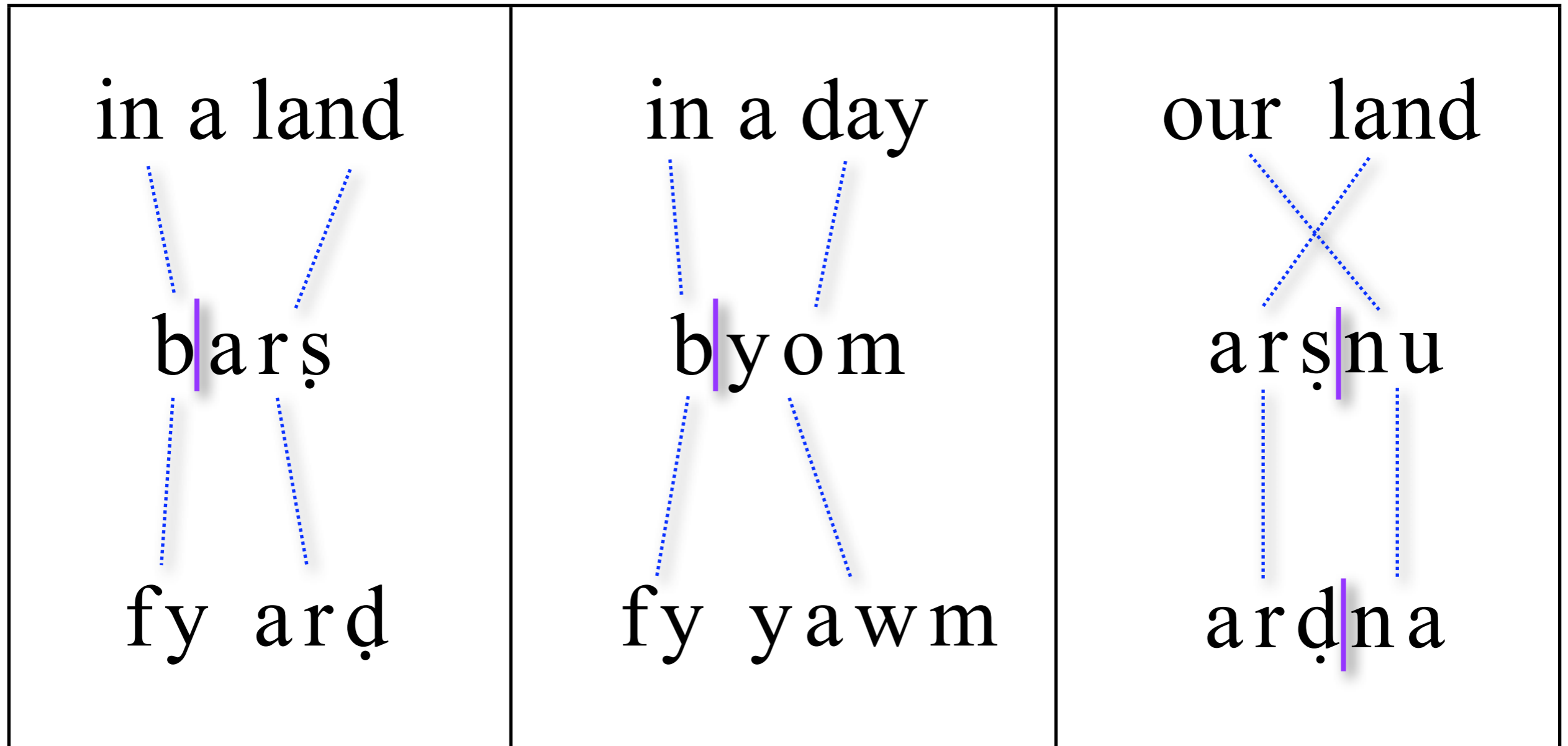
in a day

b|yom

our land

arş|nu

- Learn from different ambiguity patterns



- Learn from different ambiguity patterns
- Learn from shared structure
  - ▶ surface similarity ( $\Rightarrow$  cognates)
  - ▶ morpheme inventory

# Related Work

- **Unsupervised language learning typically formulated in monolingual framework**  
(Merialdo 1994; Klein 2005; Goldwater 2007)
- **Projection of annotations from resource-rich to resource-poor language via parallel corpora**  
(Yarowsky et al 2000; Xi & Hwa 2005; Klementiev & Roth 2006)

# A Generative Approach

Bayesian non-parametric model:

- ▶ Capture multilingual patterns:
  1. morpheme co-occurrence
  2. surface similarity (cognates)
- ▶ Allow language-specific idiosyncrasies

Challenge: Jointly model alignments and segmentations

English: A dog ate the cat

Arabic: <sup>ʔ</sup>kl klb al-qṭh

Hebrew: <sup>ʔ</sup>kl klb <sup>ʔ</sup>t h-ḥṭul

English:

A

dog

ate

the

cat

Arabic:

<sup>ʔ</sup>kl

klb

al-qṭh

Hebrew:

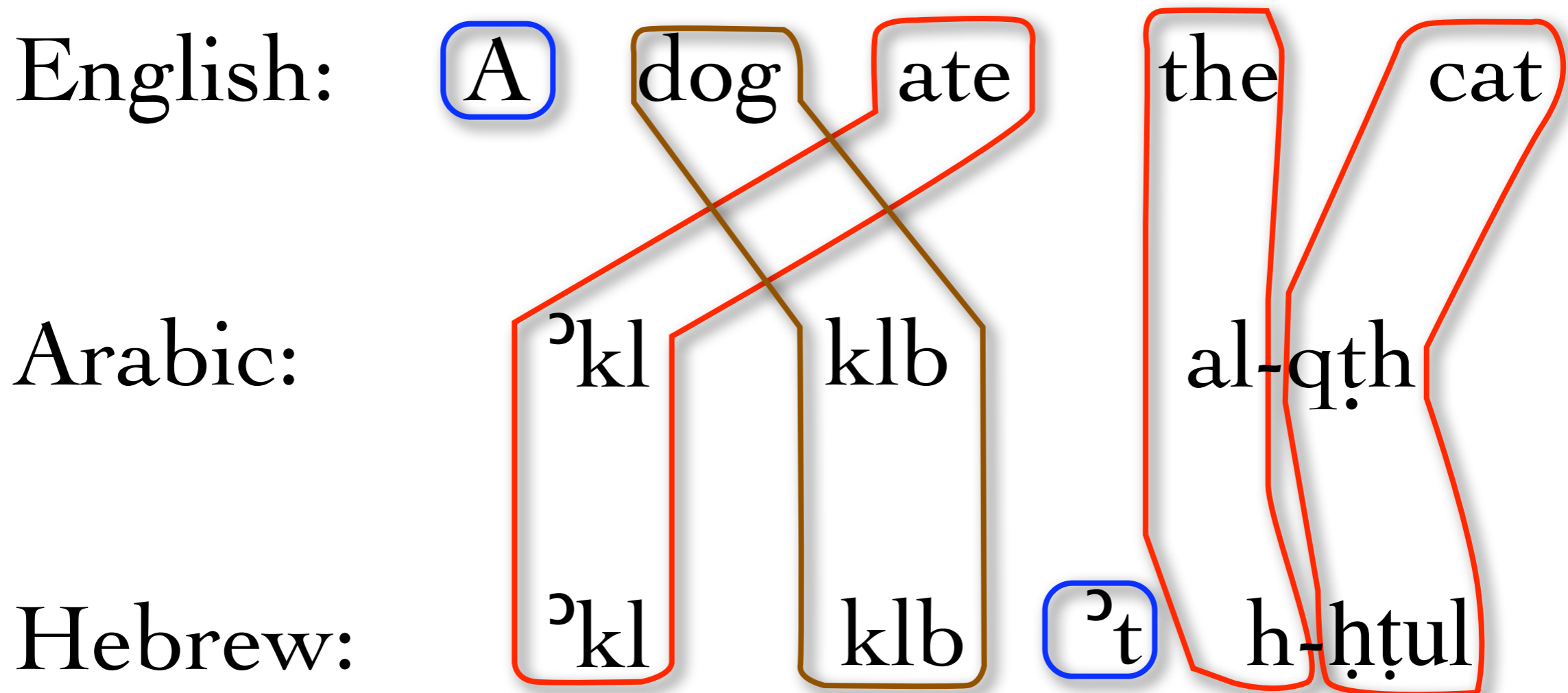
<sup>ʔ</sup>kl

klb

<sup>ʔ</sup>t

h-ḥṭul

*Coupled Morphemes:* cross-lingual morpheme tuples with common function



**Coupled Morphemes:** cross-lingual morpheme tuples with common function

**Stray Morphemes:** morphemes with no analogue in other language



# Generative Sketch

1. Draw distributions (“parameters”) from priors
2. *For each bilingual parallel phrase:*
  - (a) Choose composition of phrase  
(number of stray and coupled morphemes)
  - (b) Generate stray and coupled morphemes
  - (c) *Separately for each language:*
    - (i) Order resulting morphemes
    - (ii) Fuse ordered morphemes into words

# Generating: *other gods* in English and Hebrew

COMPOSE:

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʕlh  
god

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʔlh	ʔhr
god	other

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʔlh	ʔhr	im
god	other	s

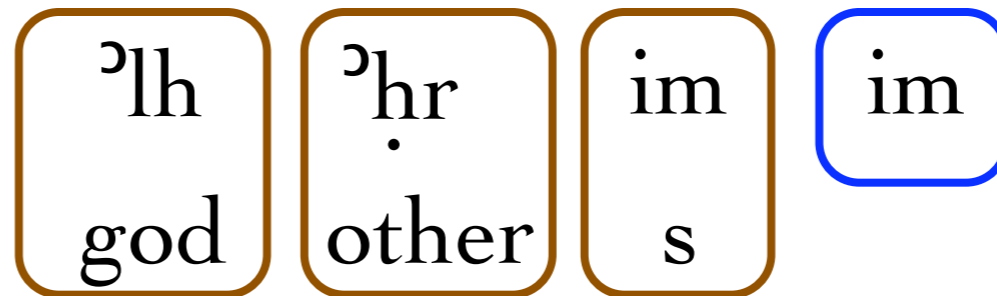
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



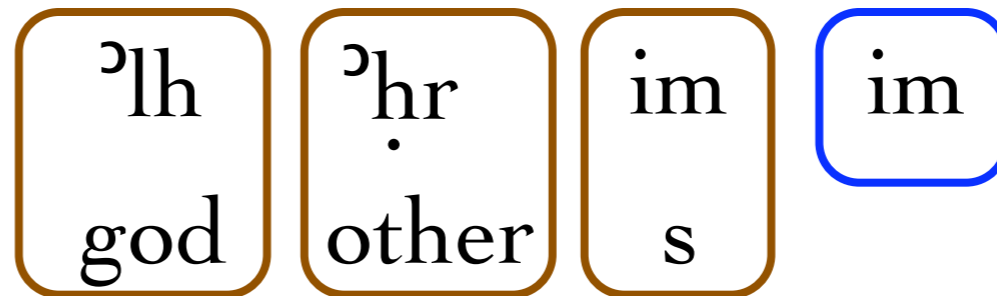


# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



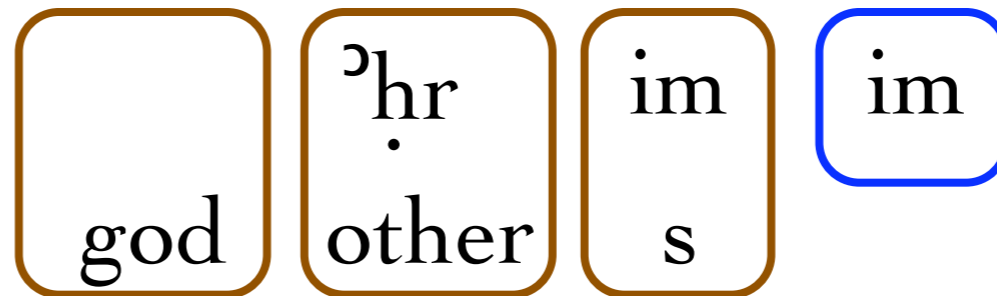
ORDER:

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

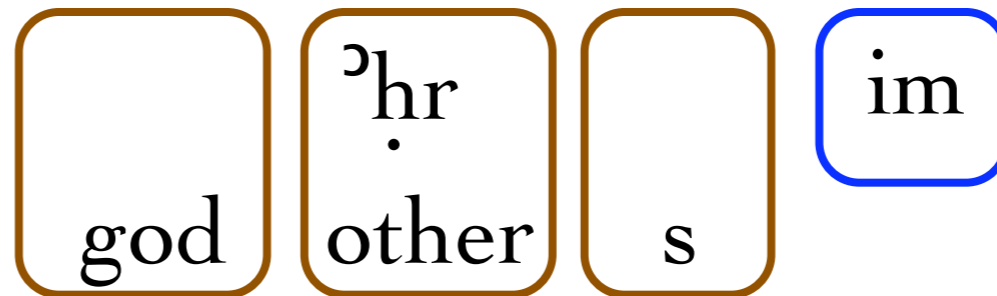
אֵלִים

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

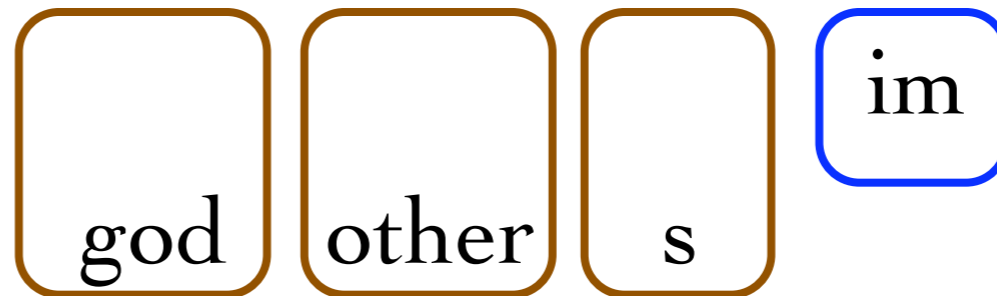
ʔh im

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

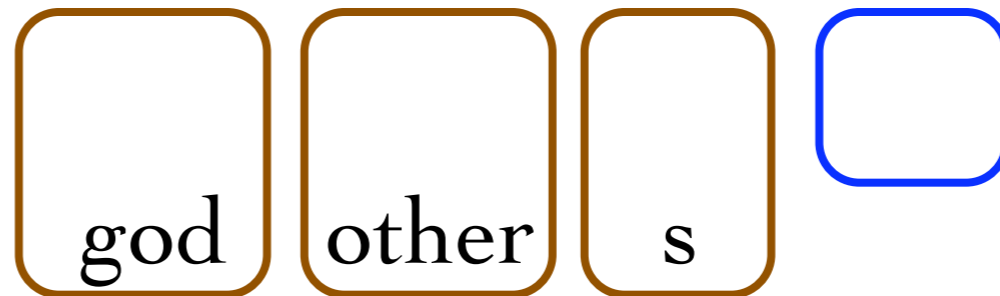
ʔlh im ʔhr

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

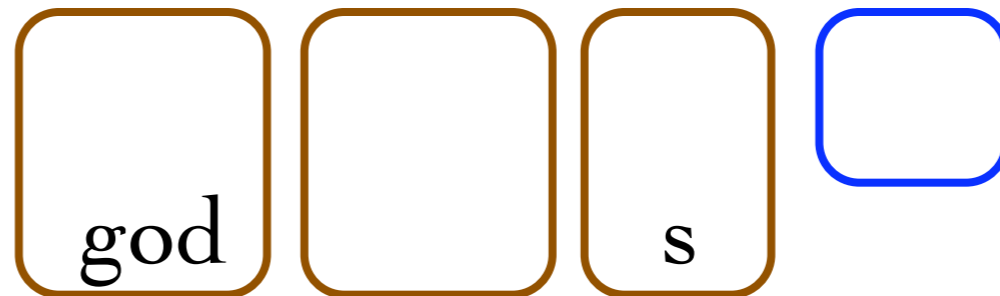
ʔlh im ʔhr im

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

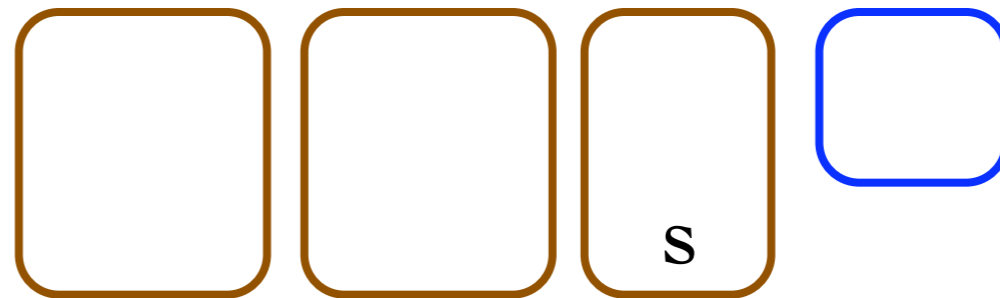
ʔlh    im    ʔhr    im  
other

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

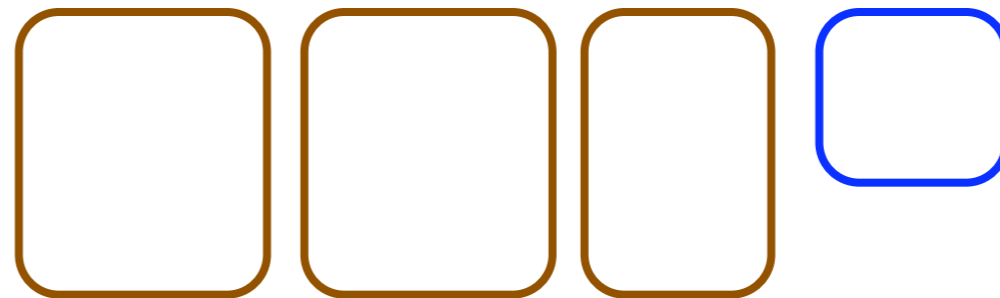
ʔlh    im    ʔhr    im  
other    god

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

ʔlh    im    ʔhr    im  
other    god    s

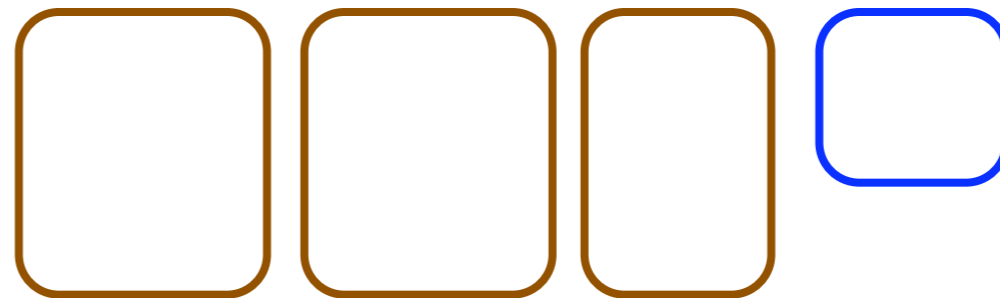


# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUZE:

ʔlh    im    ʔhr    im  
other    god    s

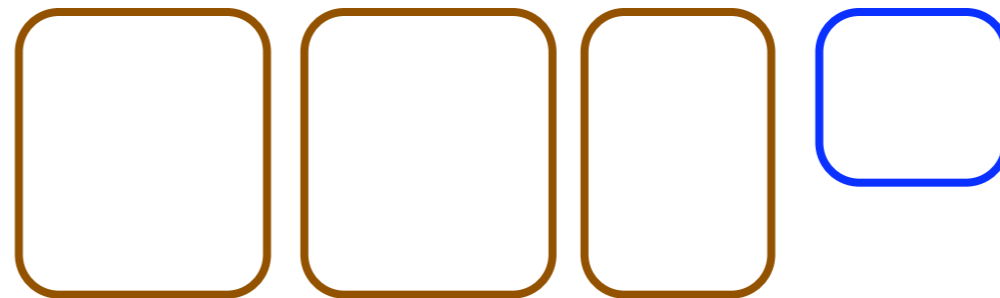
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUZE:

<sup>ʔ</sup>lhim                    <sup>ʔ</sup>hr                    im  
other    god                    s

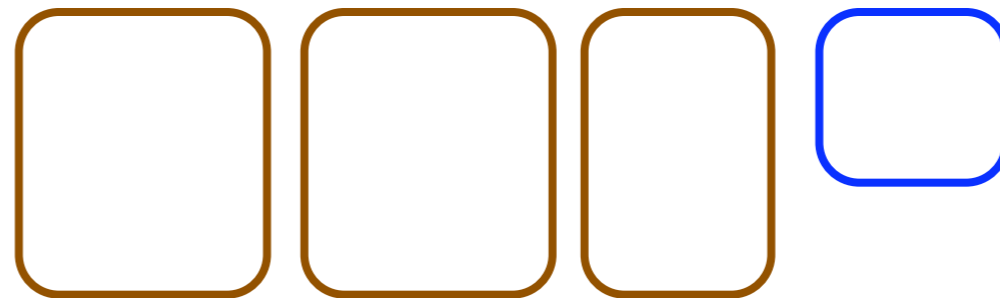
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUUSE:

ʔlhim            ʔh·rim

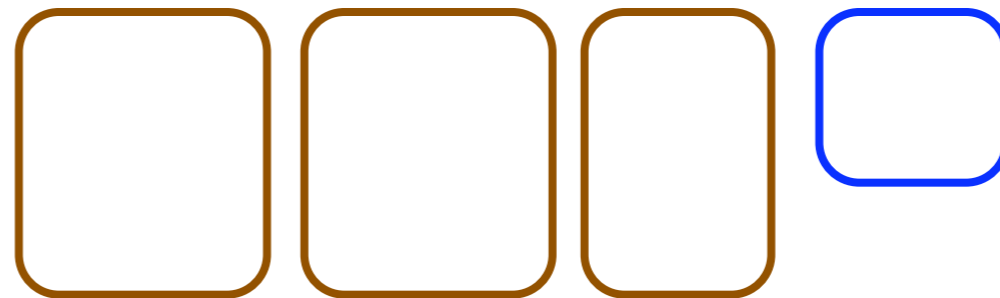
other    god        s

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUZE:

<sup>ʔ</sup>lhim      <sup>ʔ</sup>h·rim  
other    gods

# Morpheme Distributions

- Coupled morpheme distribution:

$A$

- Stray morpheme distributions for each language:

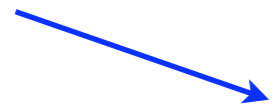
$E, F$

- Priors with favorable resulting properties:

- ▶ Infinite
- ▶ Concentrated on few patterns

# Prior on Stray Morpheme Distributions

Stray morpheme in first  
language



$e$

$\sim$

$E$

Distributions over strings in first  
language



Stray morpheme in  
second language



$f$

$\sim$

$F$

Distributions over strings in  
second language



# Prior on Stray Morpheme Distributions

Stray morpheme in first language

$$e \sim E$$

Stray morpheme in second language

$$f \sim F$$

$$E \sim DP(\alpha, P_e)$$

$$F \sim DP(\alpha, P_f)$$

Concentration Parameter

Base Distribution

# Prior on Stray Morpheme Distributions

Stray morpheme in first language

$$e \sim E$$

Stray morpheme in second language

$$f \sim F$$

$$E \sim DP(\alpha, P_e)$$

$$F \sim DP(\alpha, P_f)$$

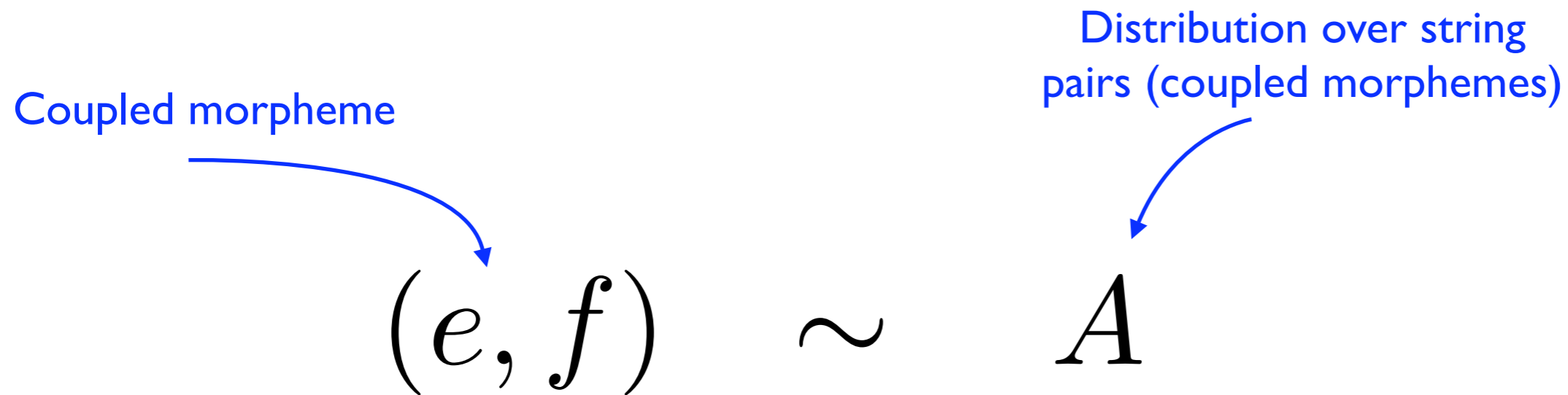
Concentration Parameter

Base Distribution

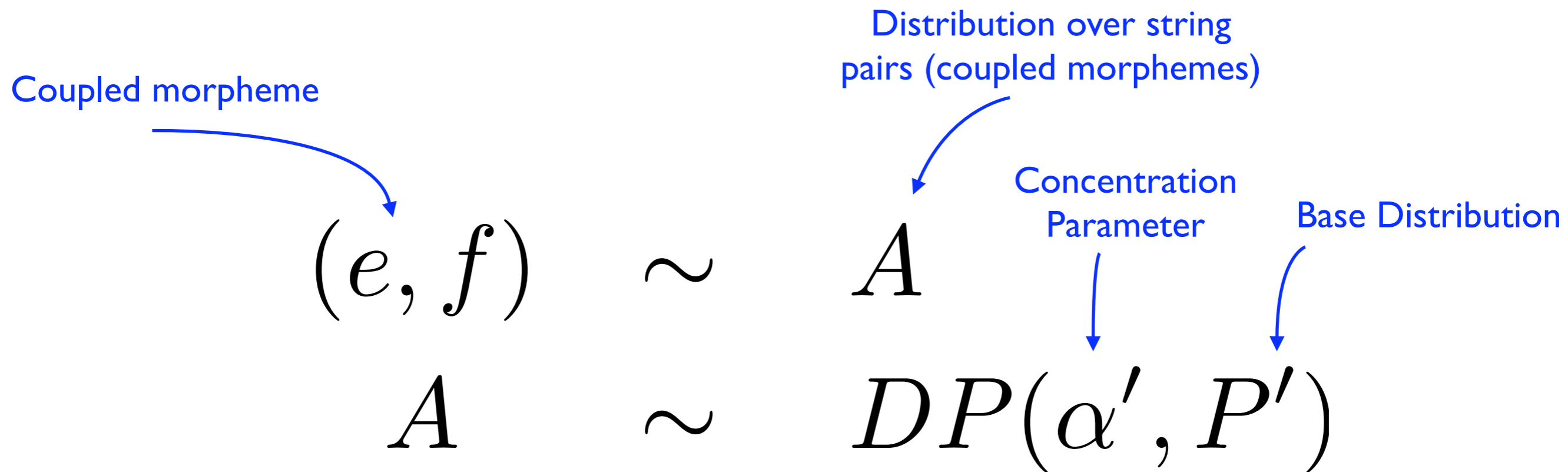
$P_e/P_f$  : geometric in length of string with special end character



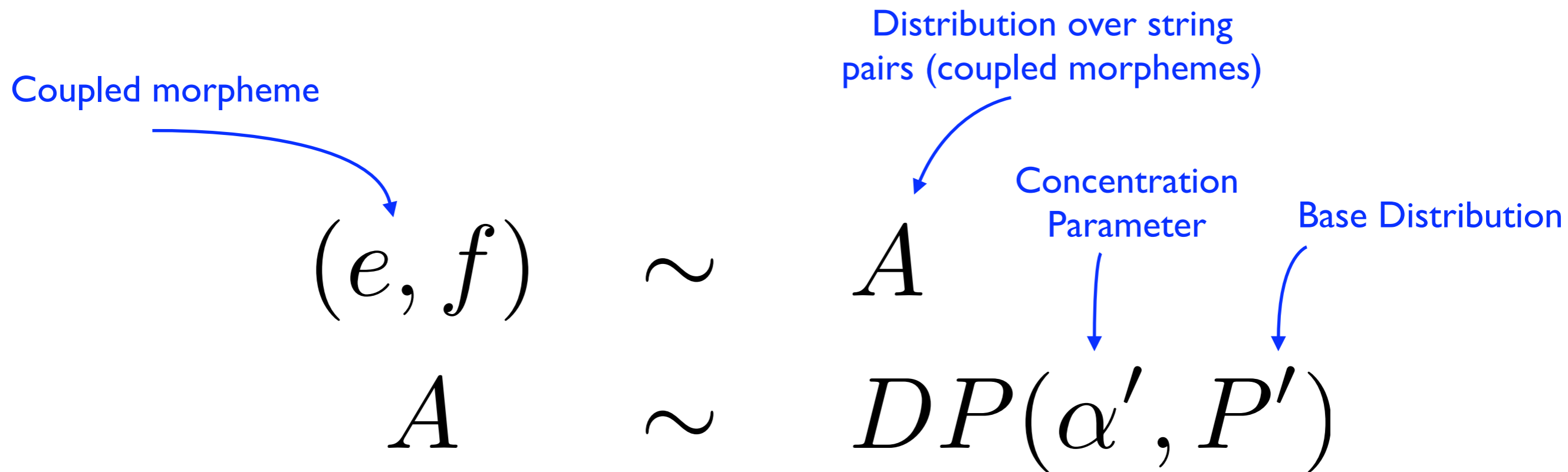
# Prior on Coupled Morpheme Distribution



# Prior on Coupled Morpheme Distribution



# Prior on Coupled Morpheme Distribution



$$P'(e, f) : ???$$

# Exploiting Surface Similarity

$$P'(e, f) = \left\{ \begin{array}{l} \text{Geometric in lengths of } e \text{ and } f \\ \text{(if languages unrelated)} \\ \\ \text{or} \\ \\ \text{Probabilistic string edit distance} \\ \text{between } e \text{ and } f \text{ (Ristad \& Yianilos 1997)} \end{array} \right.$$

# Exploiting Surface Similarity

$$P'(e, f) = \left\{ \begin{array}{l} \text{Geometric in lengths of } e \text{ and } f \\ \text{(if languages unrelated)} \\ \\ \text{or} \\ \\ \text{Probabilistic string edit distance} \\ \text{between } e \text{ and } f \text{ (Ristad \& Yianilos 1997)} \end{array} \right.$$

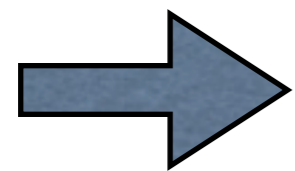
Allow substitutions between *related* letters:

$$P'(klb, klb) > P'(klb, gdy)$$

# Generative Sketch

1. Draw distributions (“parameters”) from priors

2. *For each bilingual parallel phrase:*



(a) Choose composition of phrase  
(number of stray and coupled morphemes)

(b) Generate stray and coupled morphemes

(c) *Separately for each language:*

(i) Order resulting morphemes

(ii) Fuse ordered morphemes into words

# Choose size and composition of phrase

$$m, n, k \sim \text{Poisson}(\lambda)$$

Numbers of stray  
morphemes in each  
language

Number of coupled  
morphemes

# Generating: *other gods* in English and Hebrew

COMPOSE:



# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

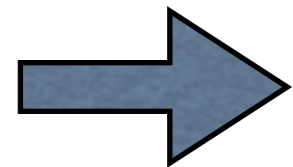
one **stray** Hebrew morpheme

# Generative Sketch

1. Draw distributions (“parameters”) from priors

2. *For each bilingual parallel phrase:*

(a) Choose composition of phrase  
(number of stray and coupled morphemes)



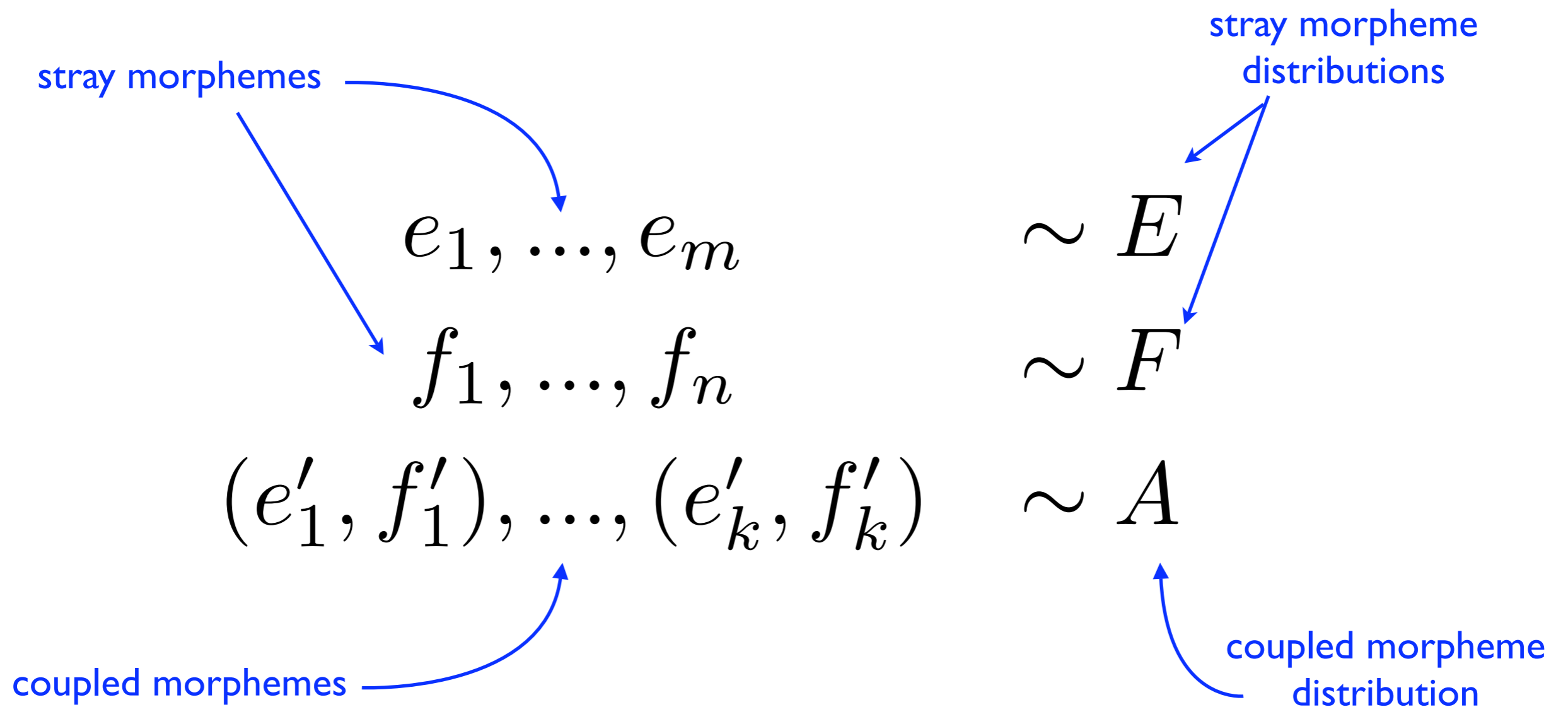
(b) Generate stray and coupled morphemes

(c) *Separately for each language:*

(i) Order resulting morphemes

(ii) Fuse ordered morphemes into words

# Generate stray and coupled morphemes



# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʕlh  
god

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʔlh	ʔhr
god	other

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:

ʔlh	ʔhr	im
god	other	s

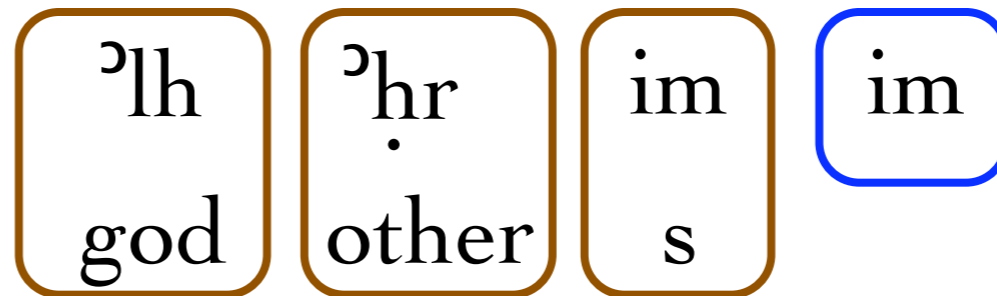
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

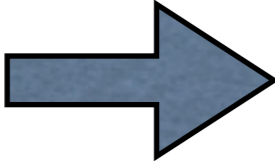
one **stray** Hebrew morpheme

GENERATE:





# Generative Sketch

1. Draw distributions (“parameters”) from priors
2. *For each bilingual parallel phrase:*
  - (a) Choose composition of phrase  
(number of stray and coupled morphemes)
  - (b) Generate stray and coupled morphemes
  - (c) *Separately for each language:*
    -  (i) Order resulting morphemes
    - (ii) Fuse ordered morphemes into words

# Order morphemes in each language

Ordered morphemes

Stray morphemes

coupled morphemes

$$\begin{array}{l} \tilde{e}_1, \dots, \tilde{e}_{m+k} \\ \tilde{f}_1, \dots, \tilde{f}_{n+k} \end{array} \sim ORDER | e_1, \dots, e_m, e'_1, \dots, e'_k$$
$$\sim ORDER | f_1, \dots, f_n, f'_1, \dots, f'_k$$

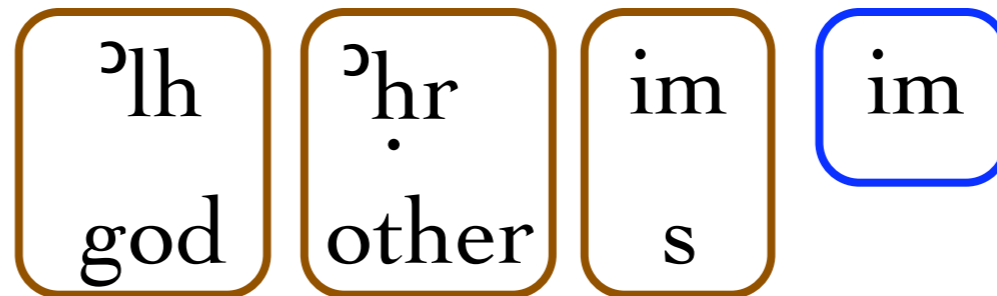
Uniform distribution over  
all possible orderings

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



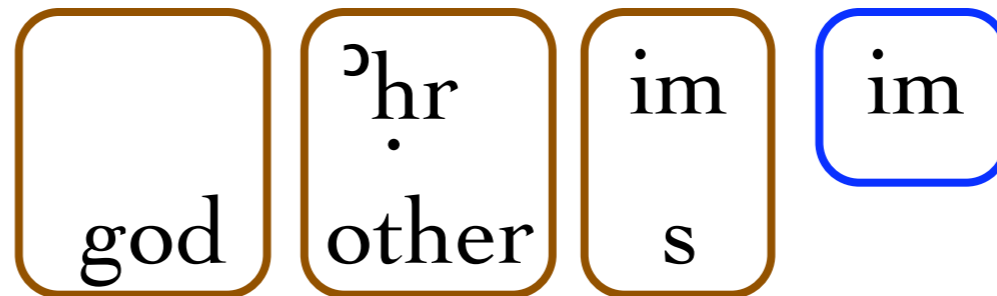
ORDER:

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

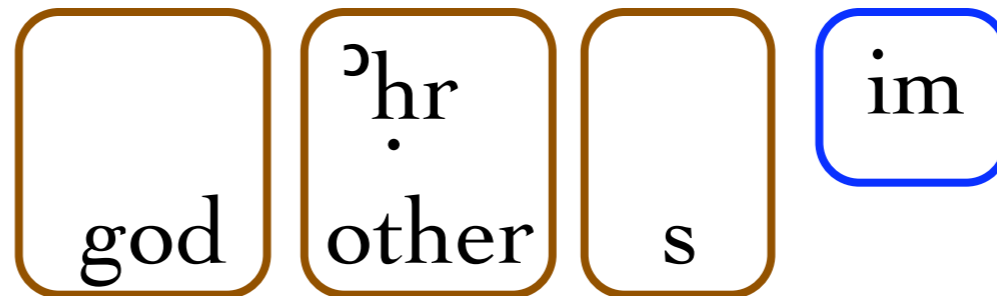
אֲחֵר

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

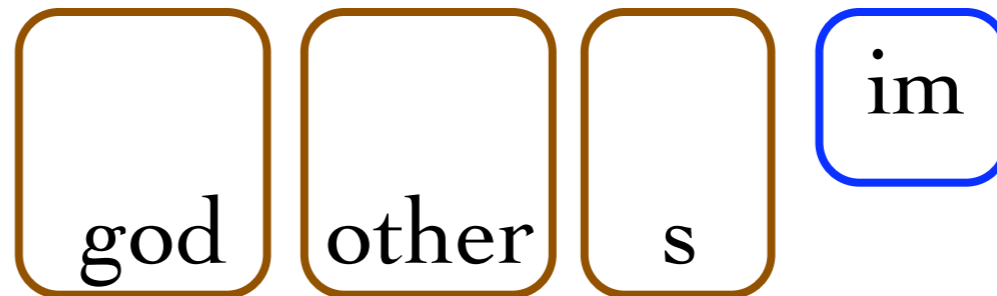
אֲחֵרִים im

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

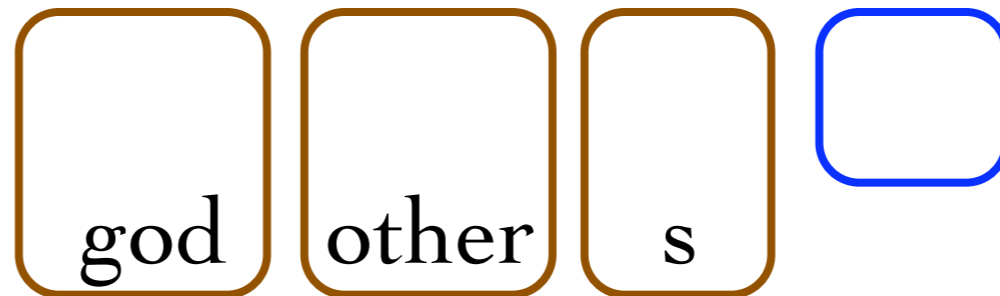
ʔlh im ʔhr

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

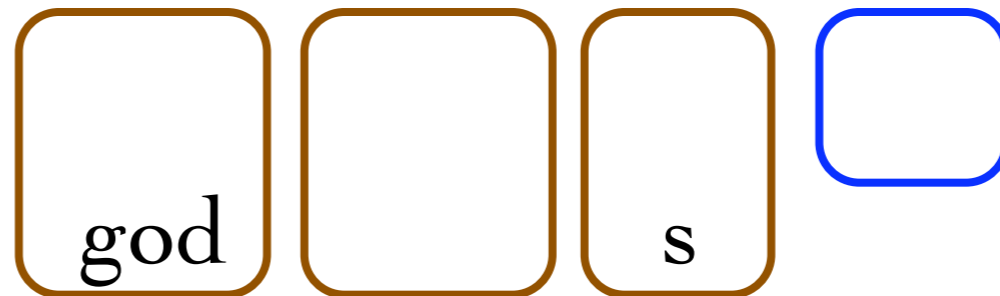
ʔlh im ʔhr im

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

ʔlh    im    ʔhr    im  
other

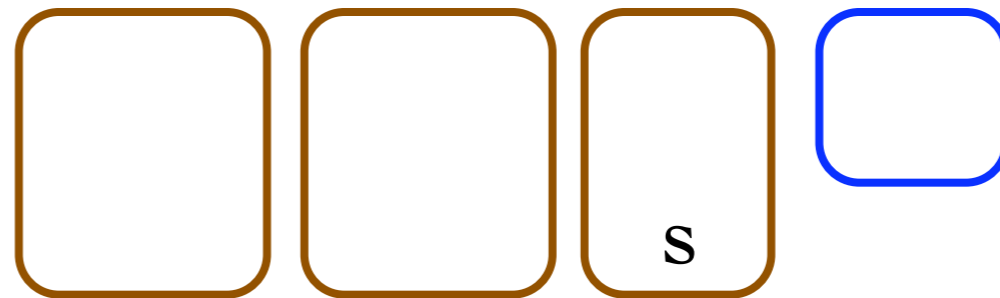


# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



ORDER:

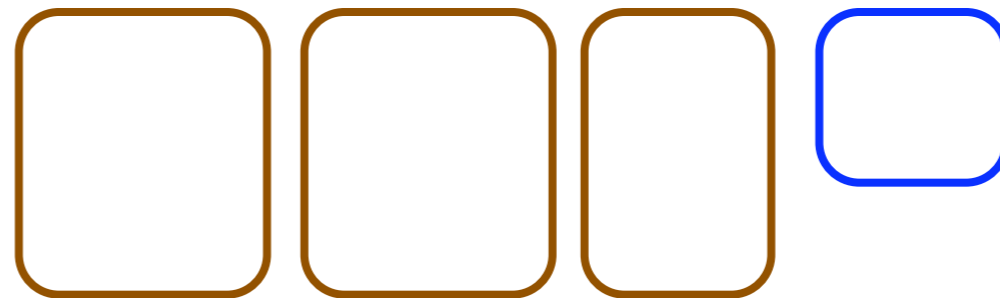
ʔlh    im    ʔhr    im  
other    god

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

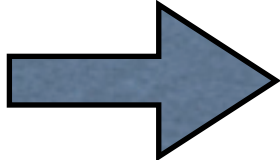
GENERATE:



ORDER:

ʔlh    im    ʔhr    im  
other    god    s

# Generative Sketch

1. Draw distributions (“parameters”) from priors
2. *For each bilingual parallel phrase:*
  - (a) Choose composition of phrase  
(number of stray and coupled morphemes)
  - (b) Generate stray and coupled morphemes
  - (c) *Separately for each language:*
    - (i) Order resulting morphemes
    -  (ii) Fuse ordered morphemes into words

# Fuse ordered morphemes into words

Phrase in first language

$$w_1, \dots, w_s \sim FUSE | \tilde{e}_1, \dots, \tilde{e}_{m+k}$$

$$v_1, \dots, v_t \sim FUSE | \tilde{f}_1, \dots, \tilde{f}_{n+k}$$

Phrase in second  
language

Ordered morphemes

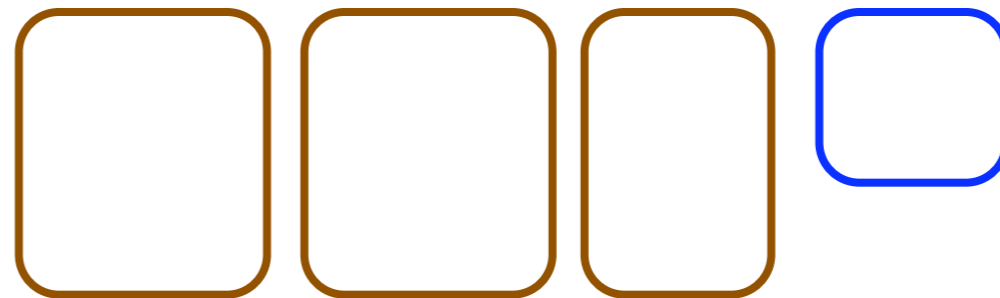
Uniform distribution  
over all possible fusings

# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme  
one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUZE:

ʔlh    im    ʔhr    im  
other    god    s

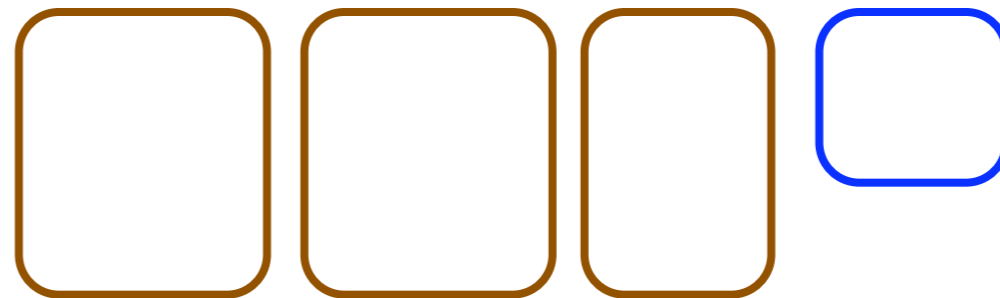
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUUSE:

<sup>ʔ</sup>lhim                    <sup>ʔ</sup>hr                    im  
other    god                    s

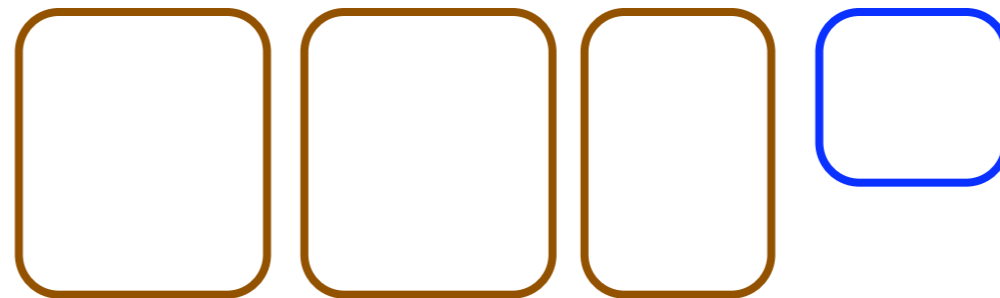
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUUSE:

ʔhim            ʔhrim

other    god        s

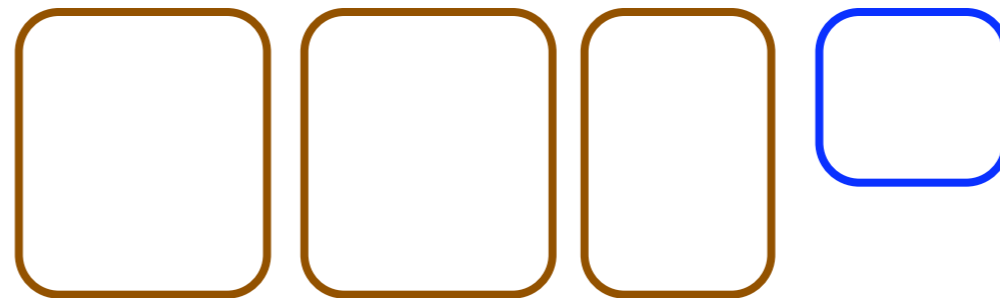
# Generating: *other gods* in English and Hebrew

COMPOSE:

three **coupled** morpheme

one **stray** Hebrew morpheme

GENERATE:



{ ORDER:  
FUZE:

ʔhim            ʔhrim

other    gods



# Gibbs Sampling

- Use standard closed forms (CRP)
- Integrate over
  - ▶ “Parameters”: coupled and stray morpheme distributions  $\theta$
  - ▶ Hidden variables: identity of stray and coupled morphemes in each phrase  $z$
- Produce segmentations with highest posterior likelihood:

$$\operatorname{argmax}_y P(y|x) = \operatorname{argmax}_y \int_{\theta} \int_z P(y, x, z, \theta) dz d\theta$$

# Data

- Bible: Hebrew, Aramaic, Arabic, English
- 6,139 short parallel phrases automatically extracted
- Words per phrase: 1 - 6
- Gold Standard:
  - ▶ Hebrew: Westminster Theological Seminary
  - ▶ Arabic: MADA (Buckwalter; Habash & Rambow 2005)
  - ▶ Morphemes per word: 1.8

# Evaluation

- Evaluate on separate *Monolingual* test data
- Examine all possible segmentation points:

בראשית  
↑↑↑↑↑

- Calculate precision, recall, and F-measure

Results for Arabic  
(similar for Hebrew)

# Morphological Segmentation: Baselines

	F-Measure
Random	18.8
Morfessor (Creutz & Lagus '05)	65.4
Monolingual Baseline	63.2

## Morfessor:

State-of-the-art unsupervised (monolingual) model

## Monolingual Baseline:

Train a single language with stray morpheme distribution  
(similar to (Goldwater 2007))

	F-Measure
Monolingual Baseline	63.2
Arabic+Hebrew	
Arabic+Aramaic	
Arabic+English	
Arabic+Aramaic (String-edit)	
Arabic+Hebrew (String-edit)	

	F-Measure
Monolingual Baseline	63.2
Arabic+Hebrew	68.3
Arabic+Aramaic	68.6
Arabic+English	
Arabic+Aramaic (String-edit)	
Arabic+Hebrew (String-edit)	

- ▶ Multilingual learning effective

	F-Measure
Monolingual Baseline	63.2
Arabic+Hebrew	68.3
Arabic+Aramaic	68.6
Arabic+English	68.5
Arabic+Aramaic (String-edit)	
Arabic+Hebrew (String-edit)	

- ▶ Multilingual learning effective
- ▶ English equally effective

	F-Measure
Monolingual Baseline	63.2
Arabic+Hebrew	68.3
Arabic+Aramaic	68.6
Arabic+English	68.5
Arabic+Aramaic (String-edit)	70.8
Arabic+Hebrew (String-edit)	72.2

- ▶ Multilingual learning effective
- ▶ English equally effective
- ▶ Best result with related languages using string-edit prior (cognates)



# Conclusions

Multilingual Learning: a new approach to unsupervised learning

- Simultaneously induce linguistic structure with interlingual links
- Exploit different patterns of ambiguity while finding shared structure

## Key findings:

- Error reduction up to 24% for morphological segmentation
- Best results: learning within language family with string-edit prior (cognates)

## Future work:

- Apply to *many* languages simultaneously
- Apply to other core NLP tasks



# Question 1.

# Question 1.

Can we exploit cross-lingual patterns to improve unsupervised language learning?

# Question 1.

Can we exploit cross-lingual patterns to improve unsupervised language learning?

- ✓ morphological segmentation  
error reduction of 24%

# Question 2.

# Question 2.

Will joint analysis provide more or less benefit when the languages belong to the same family?



# Question 2.

Will joint analysis provide more or less benefit when the languages belong to the same family?

- Unrelated but “easier” languages provide large benefit.

# Question 2.

Will joint analysis provide more or less benefit when the languages belong to the same family?

- Unrelated but “easier” languages provide large benefit.
- Best performance: within-family learning when surface similarity used.

# Thank You!

For supervised results see:

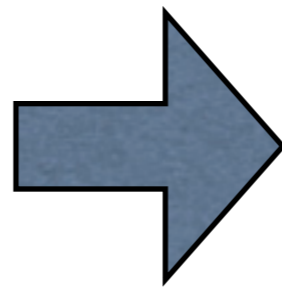
Benjamin Snyder & Regina Barzilay. *Cross-lingual Propagation for Morphological Analysis*. AAI 2008

<http://people.csail.mit.edu/bsnyder>

# Morphological Segmentation

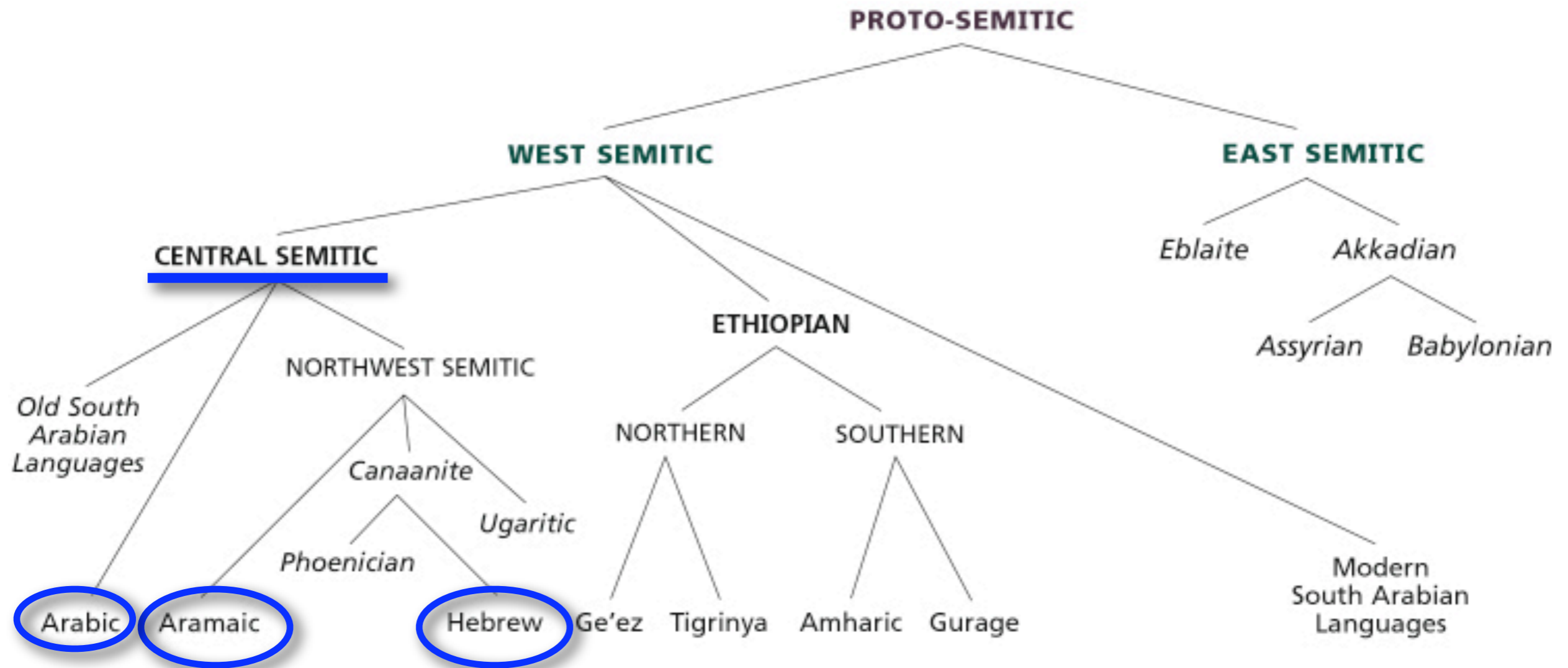
“In my land”

- Hebrew: *barʁy*
- Arabic: *fy arḁy*
- Aramaic: *barʁy*



- “In”    “land”    “my”
- *b - arʁ - y*
  - *fy arḁ - y*
  - *b - arʁ - y*

# Semitic Languages



# Related Work

- Unsupervised language learning typically formulated in monolingual framework  
(Merialdo 1994; Klein 2005; Goldwater 2007)
- Projection of annotations from resource-rich to resource-poor language via parallel corpora  
(Yarowsky et al 2000; Xi & Hwa 2005; Klementiev & Roth 2006)

# Generative Outline

- Draw parameters from priors:  $\theta \sim P(\theta)$
- Generate hidden structure from parameters:  $z \sim P(z|\theta)$
- Produce bilingual parallel phrases from hidden structure:  $x \sim P(x|z)$

“Read off” segmentation  $x, z \Rightarrow y$

# Unsupervised NLP?

In the beginning God created the heavens and the earth.

בראשית ברא אלהים את השמים ואת הארץ

Large performance gap



# Unsupervised NLP?

In the beginning God created the heavens and the earth.

בראשית ברא אלהים את השמים ואת הארץ

Large performance gap

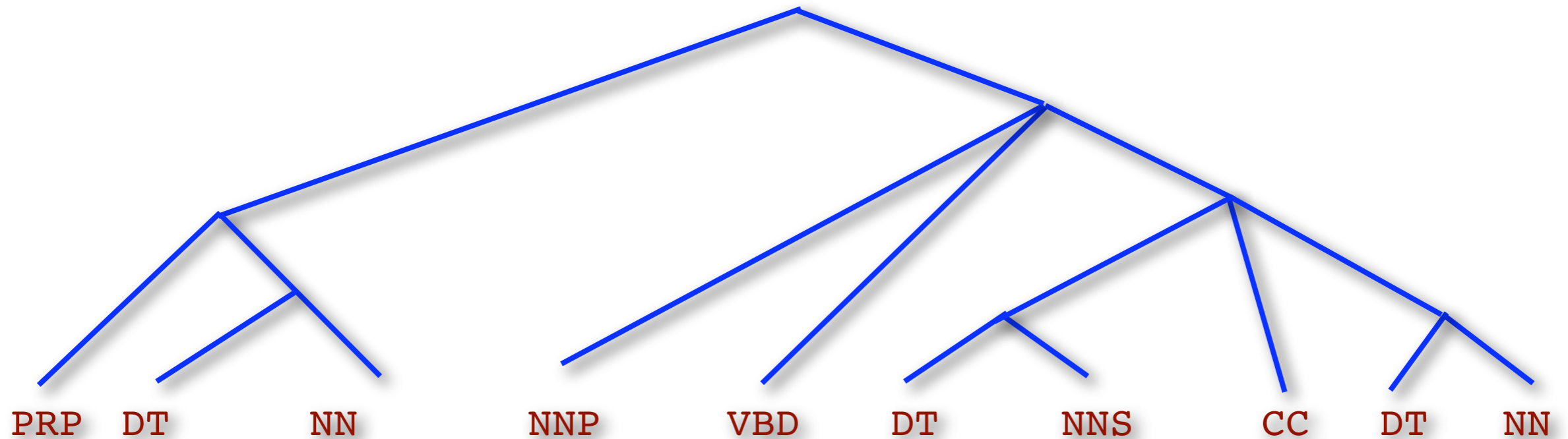
# Unsupervised NLP?

PRP DT NN NNP VBD DT NNS CC DT NN  
In the beginning God created the heavens and the earth.

בראשית ברא אלהים את השמים ואת הארץ

Large performance gap

# Unsupervised NLP?



In the begin|ning God create|d the heaven|s and the earth.

בראשית ברא אלהים את השמים ואת הארץ

Large performance gap

# Multilingual Corpora: A Rosetta Stone for unsupervised NLP

Semitic  
Languages

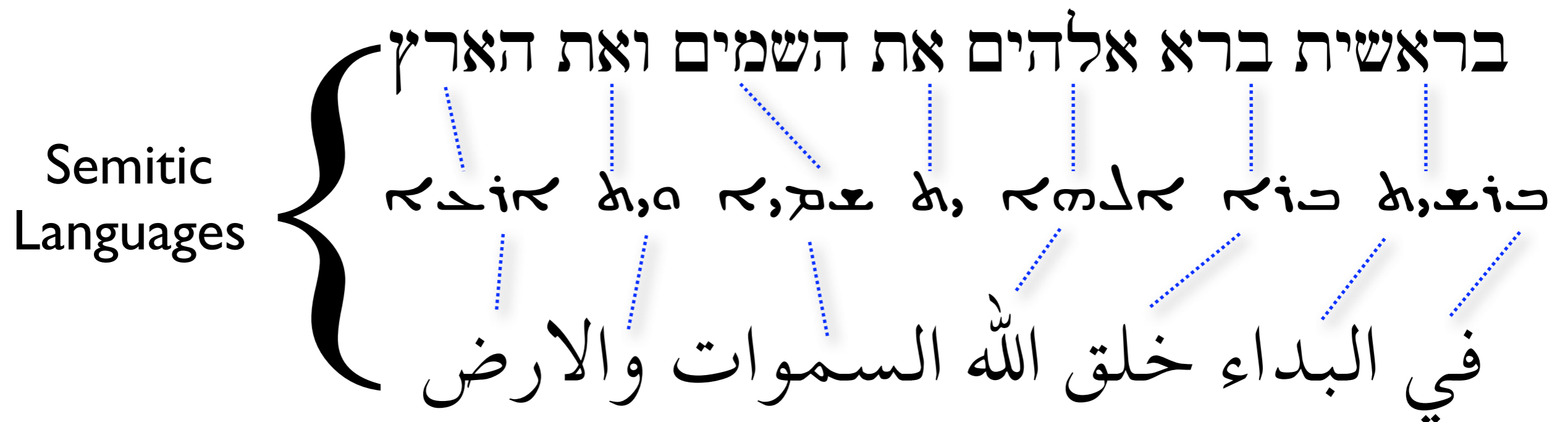
בראשית ברא אלהים את השמים ואת הארץ

בִּרְאֵת בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والارض

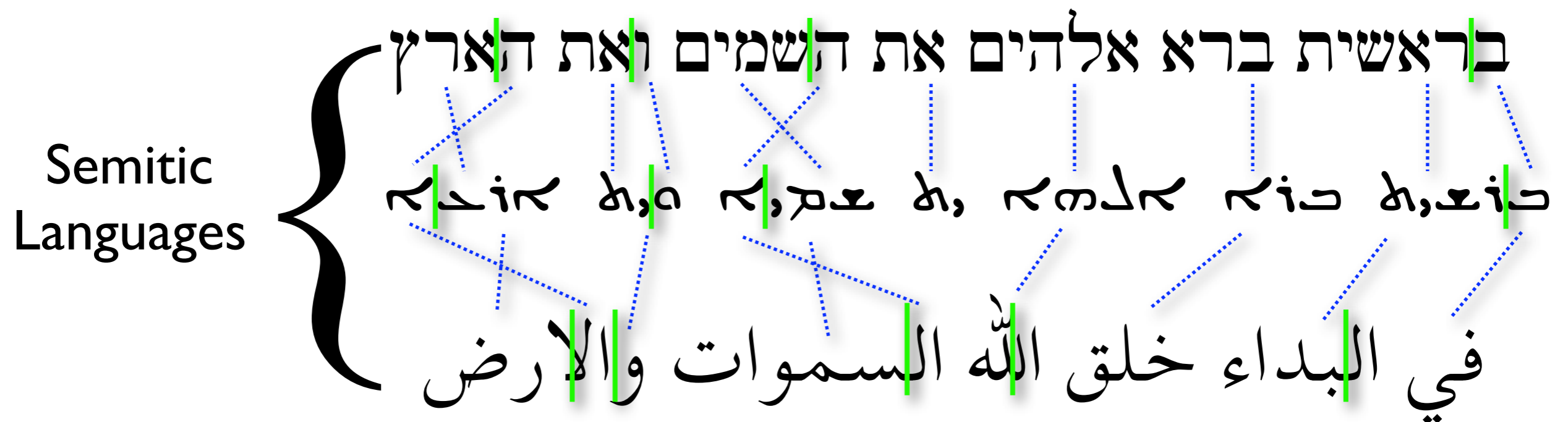
“In the beginning God created the heavens and the earth...”

# Multilingual Corpora: A Rosetta Stone for unsupervised NLP

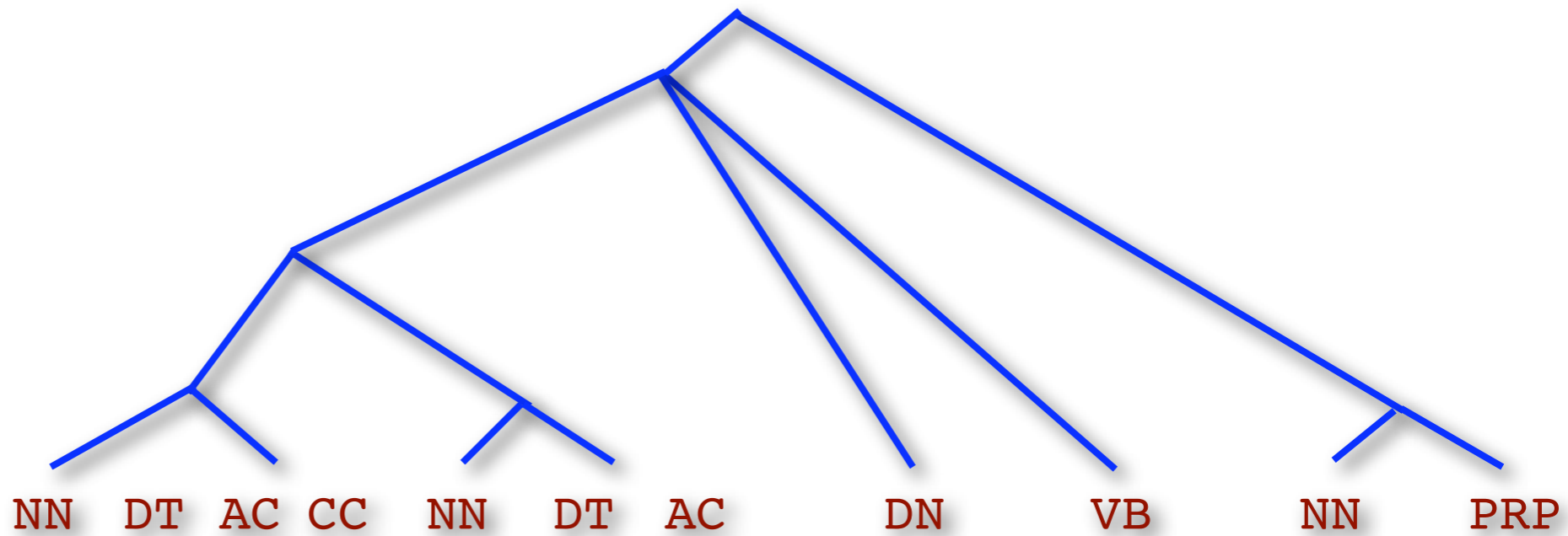


“In the beginning God created the heavens and the earth...”

# Multilingual Corpora: A Rosetta Stone for unsupervised NLP



“In the beginning God created the heavens and the earth...”

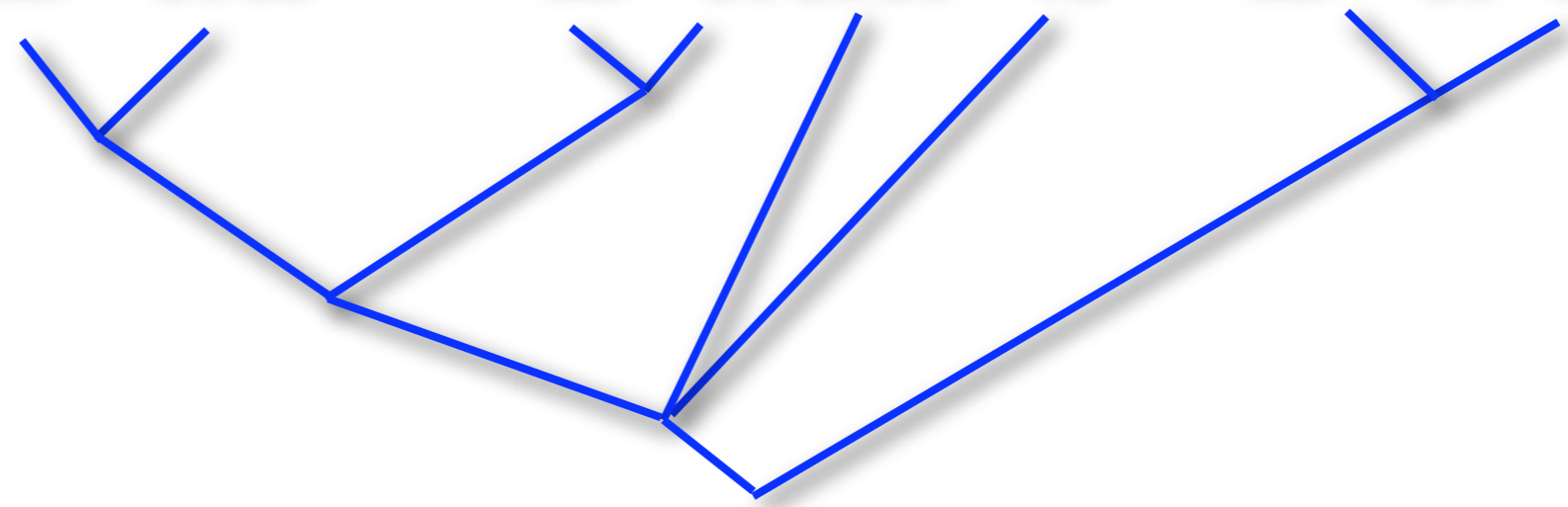


בראשית ברא אלהים את השמים ואת הארץ

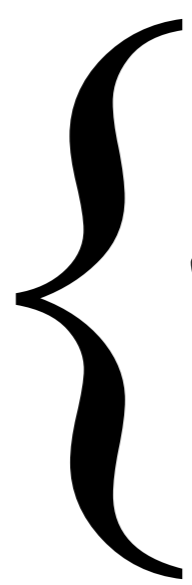
בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP



Semitic Languages



Semitic  
Languages

בראשית ברא אלהים את השמים ואת הארץ

בְּרֵאשִׁית בְּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ

في البدء خلق الله السموات والأرض