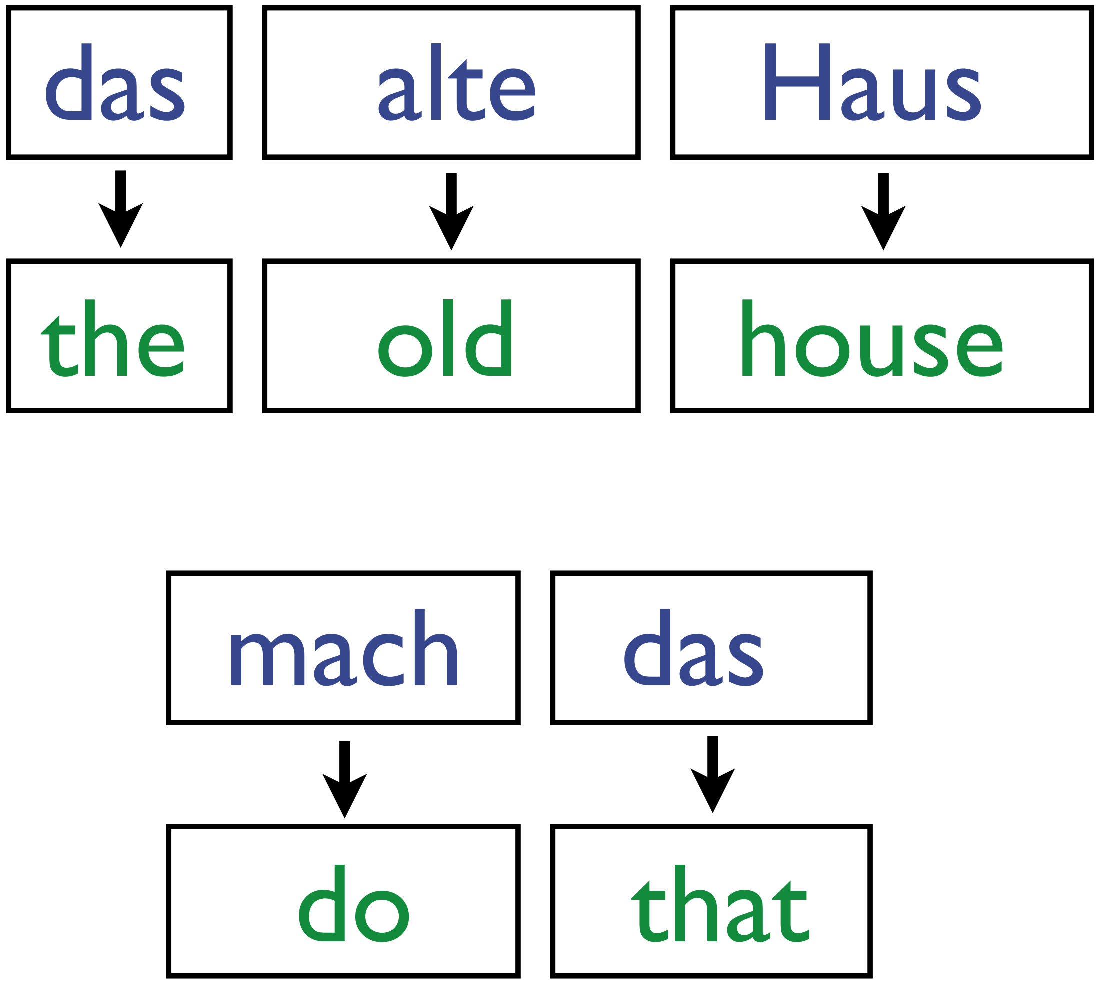
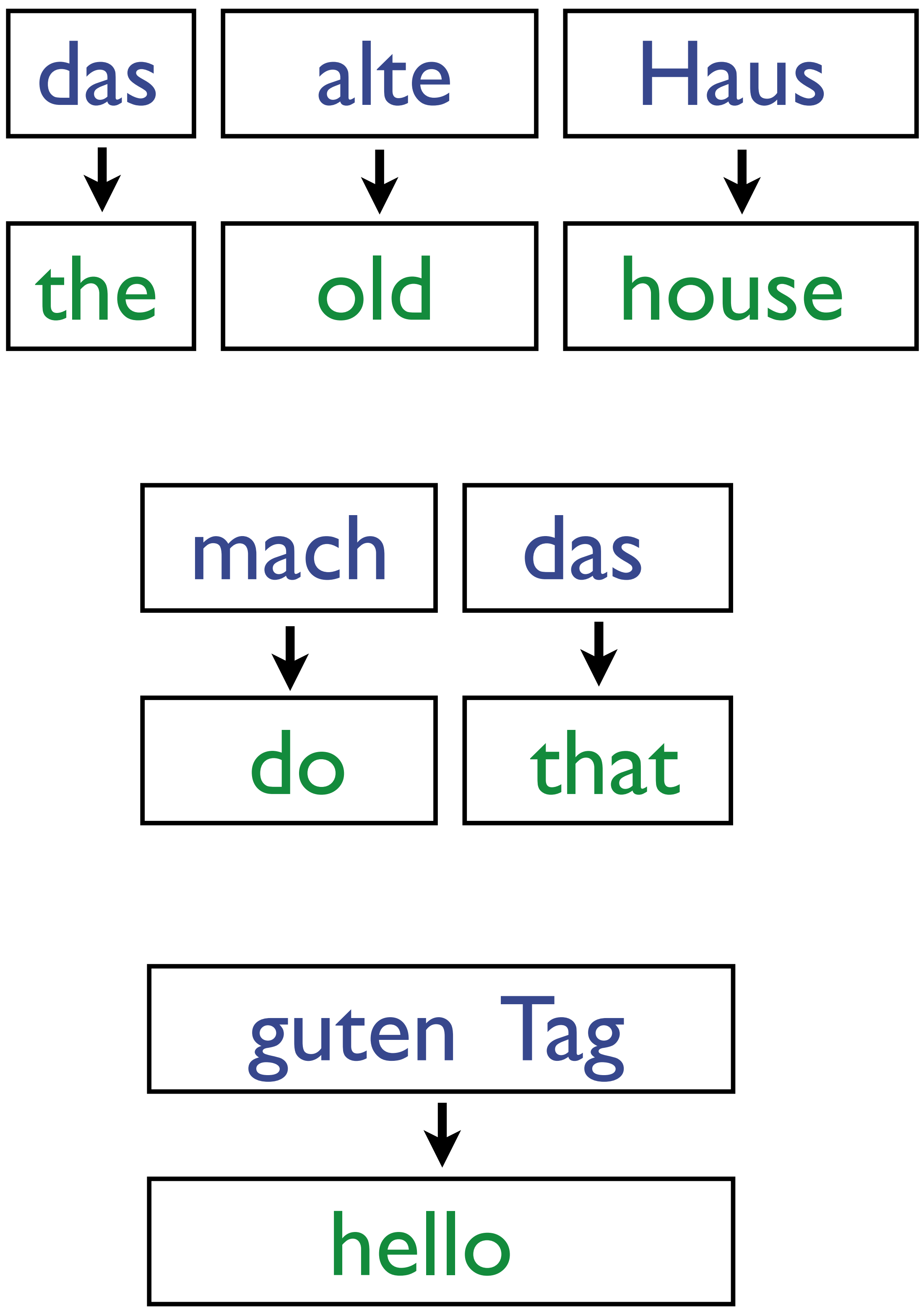
Integrating Morphology in Probabilistic Translation Models

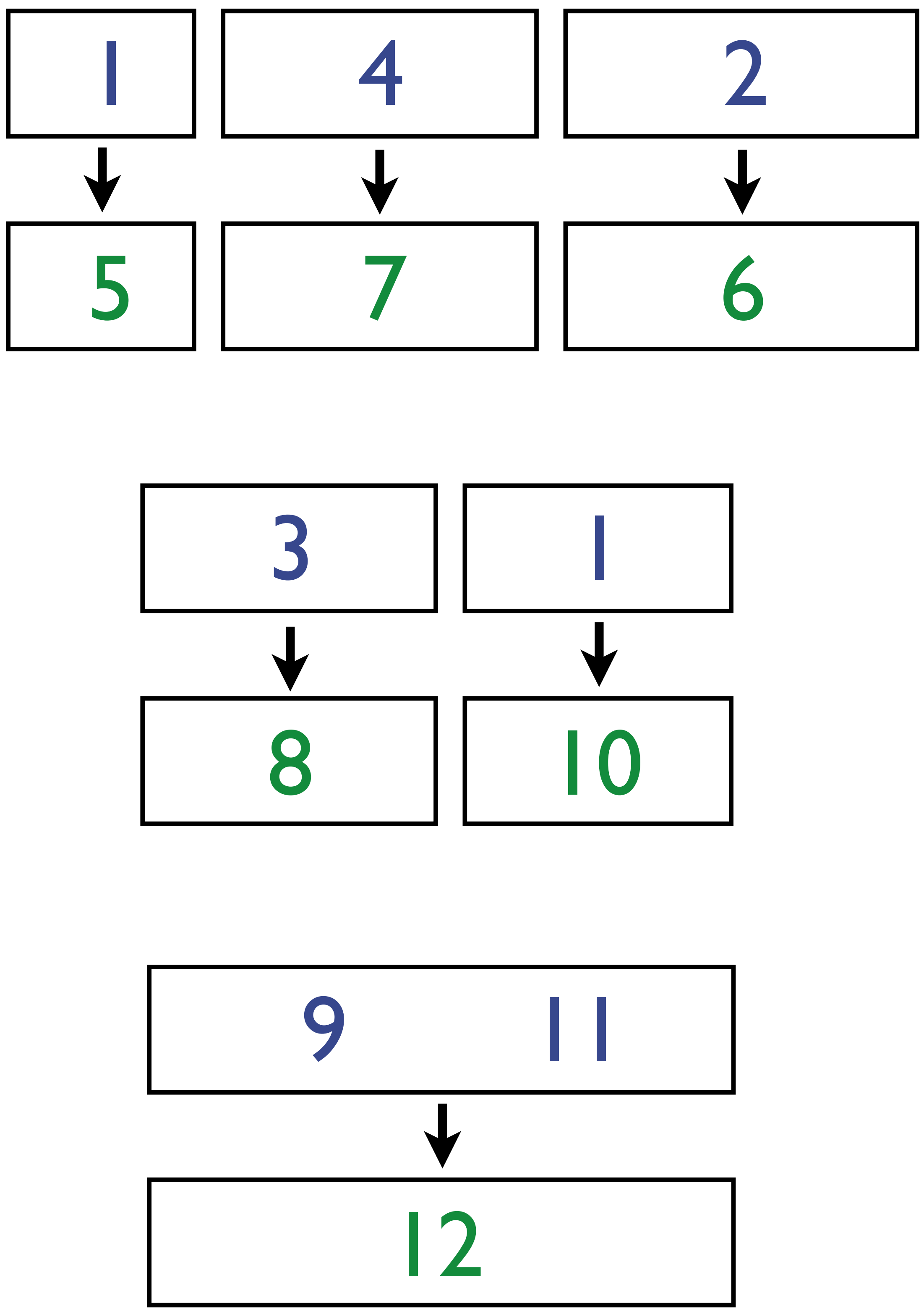
Chris Dyer

joint work with Jon Clark, Alon Lavie, and Noah Smith

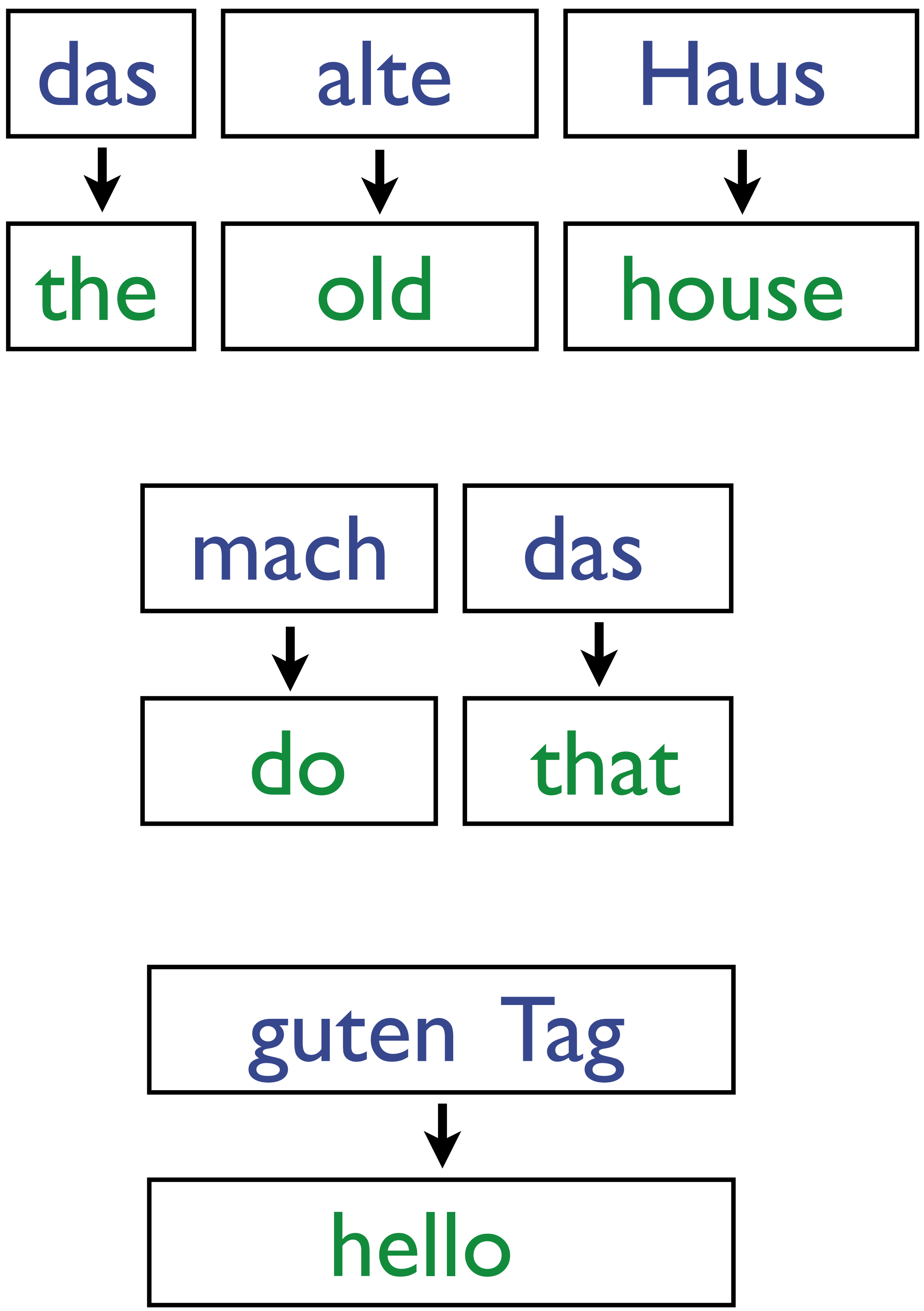
January 24, 2011

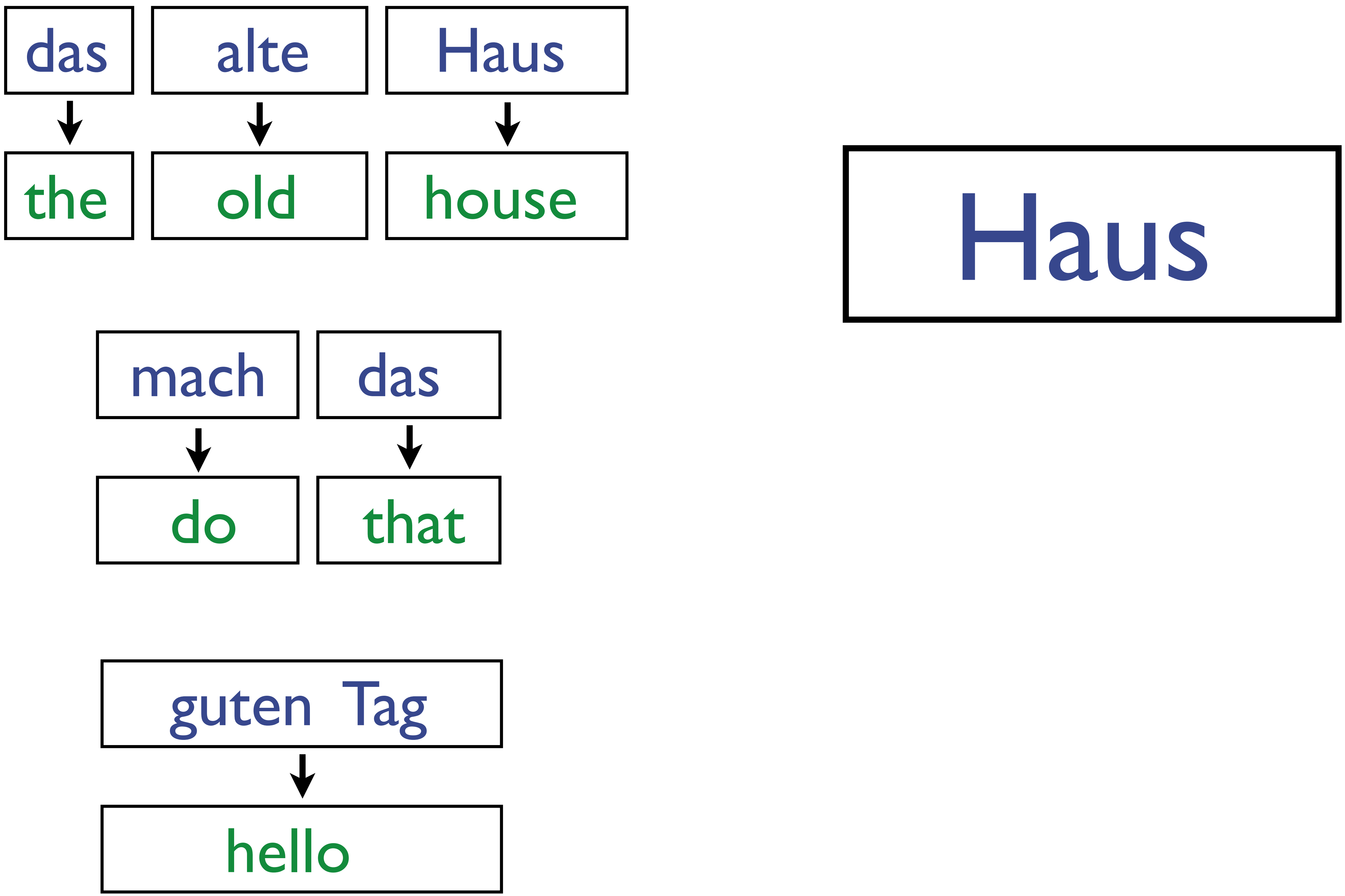

das
alte
Haus
the
old
house
mach
das
do
that

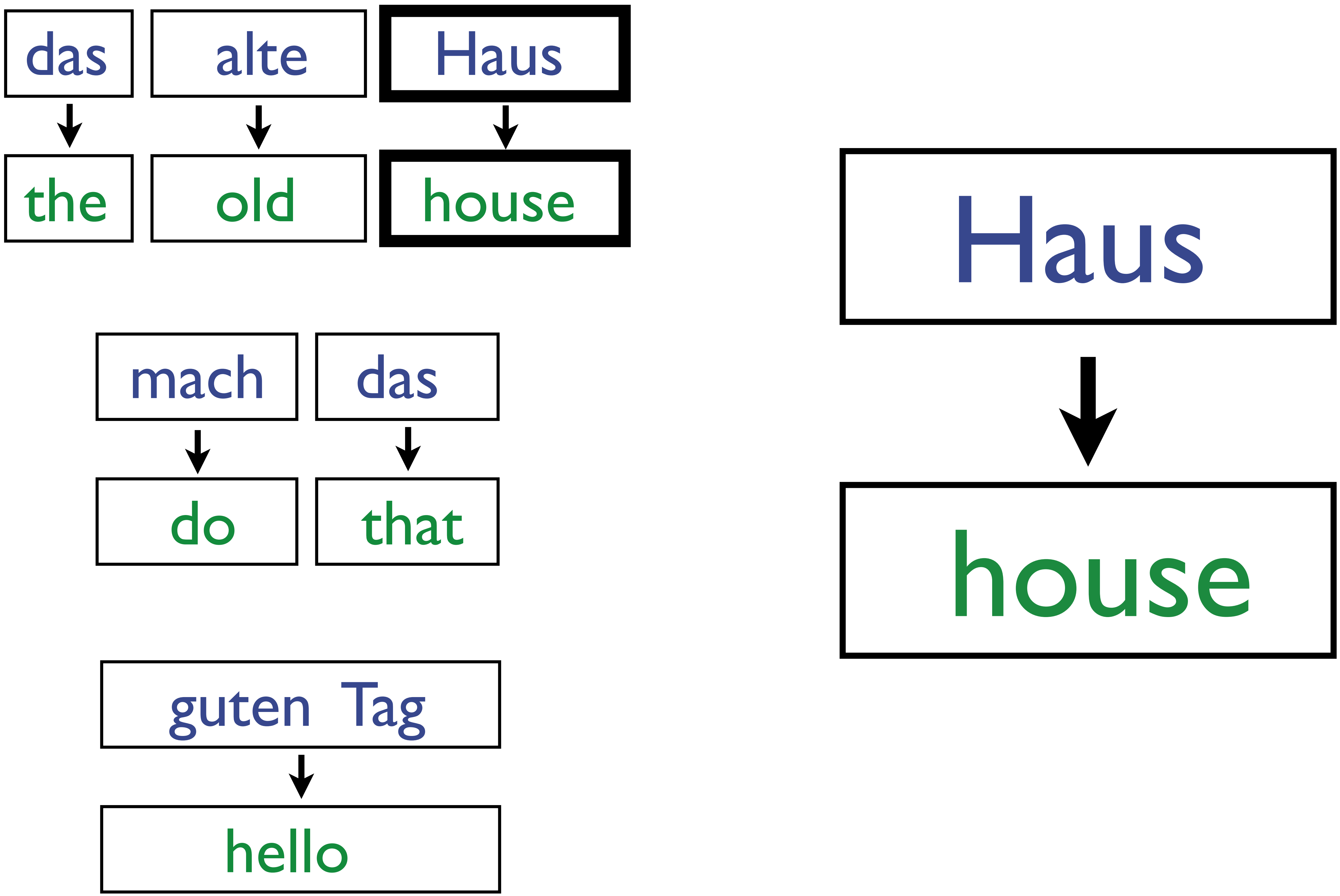

das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello

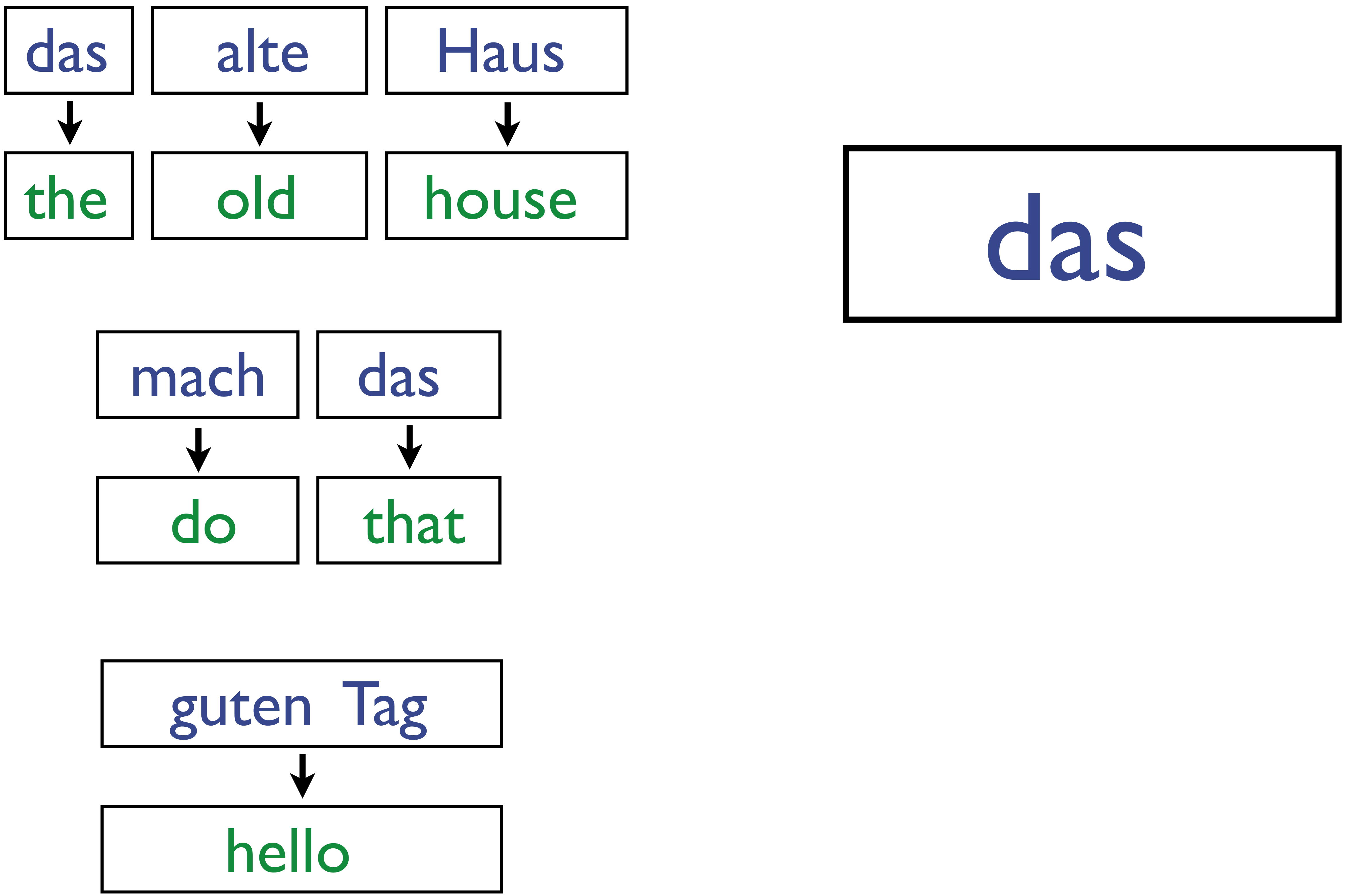


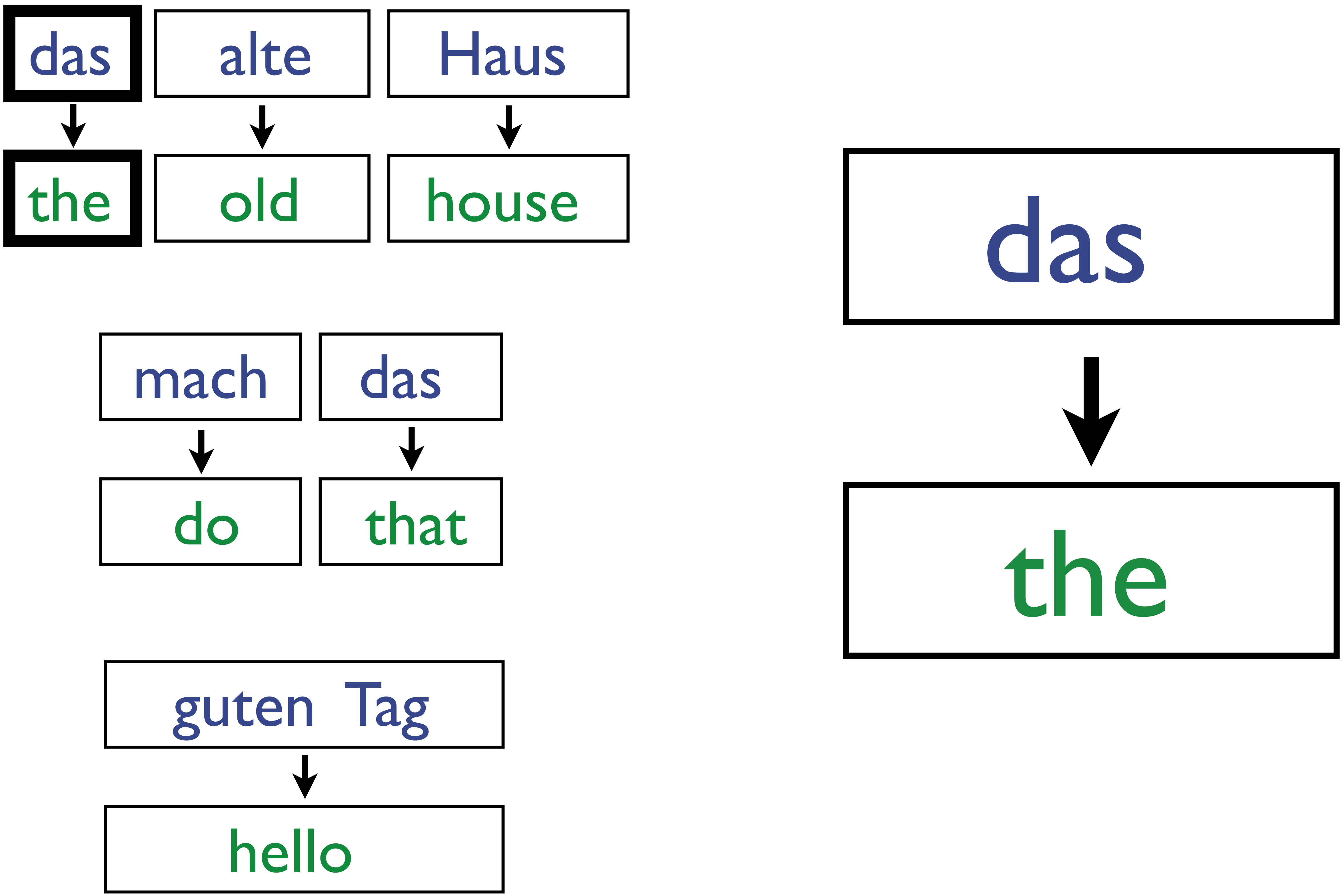
1
4
2
5
7
6
3
1
8
10
9
11
12

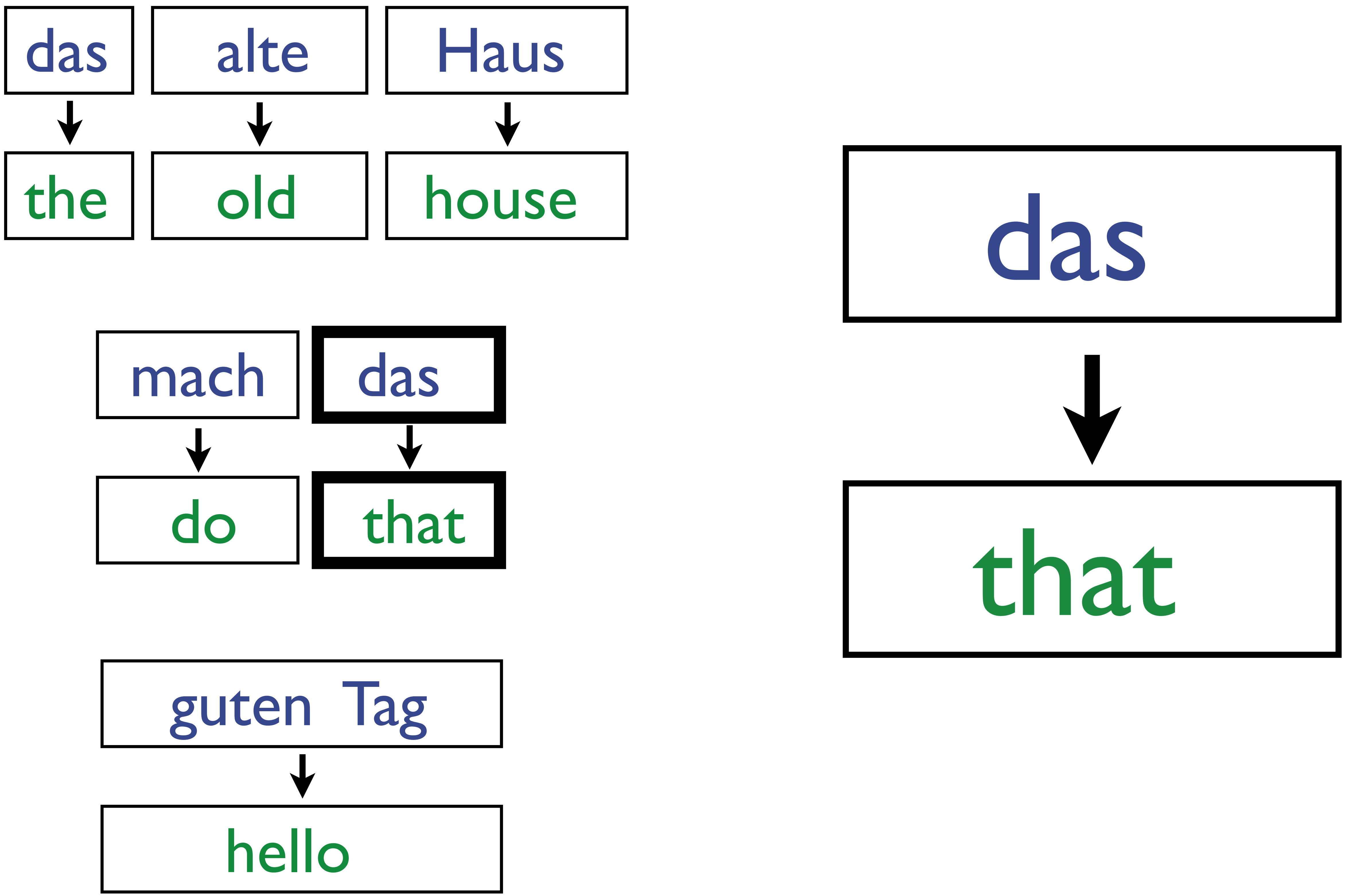

das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
Haus


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
Haus
house


das
alte
Haus
the
old
house
das
mach
das
do
that
guten Tag
hello


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
das
the

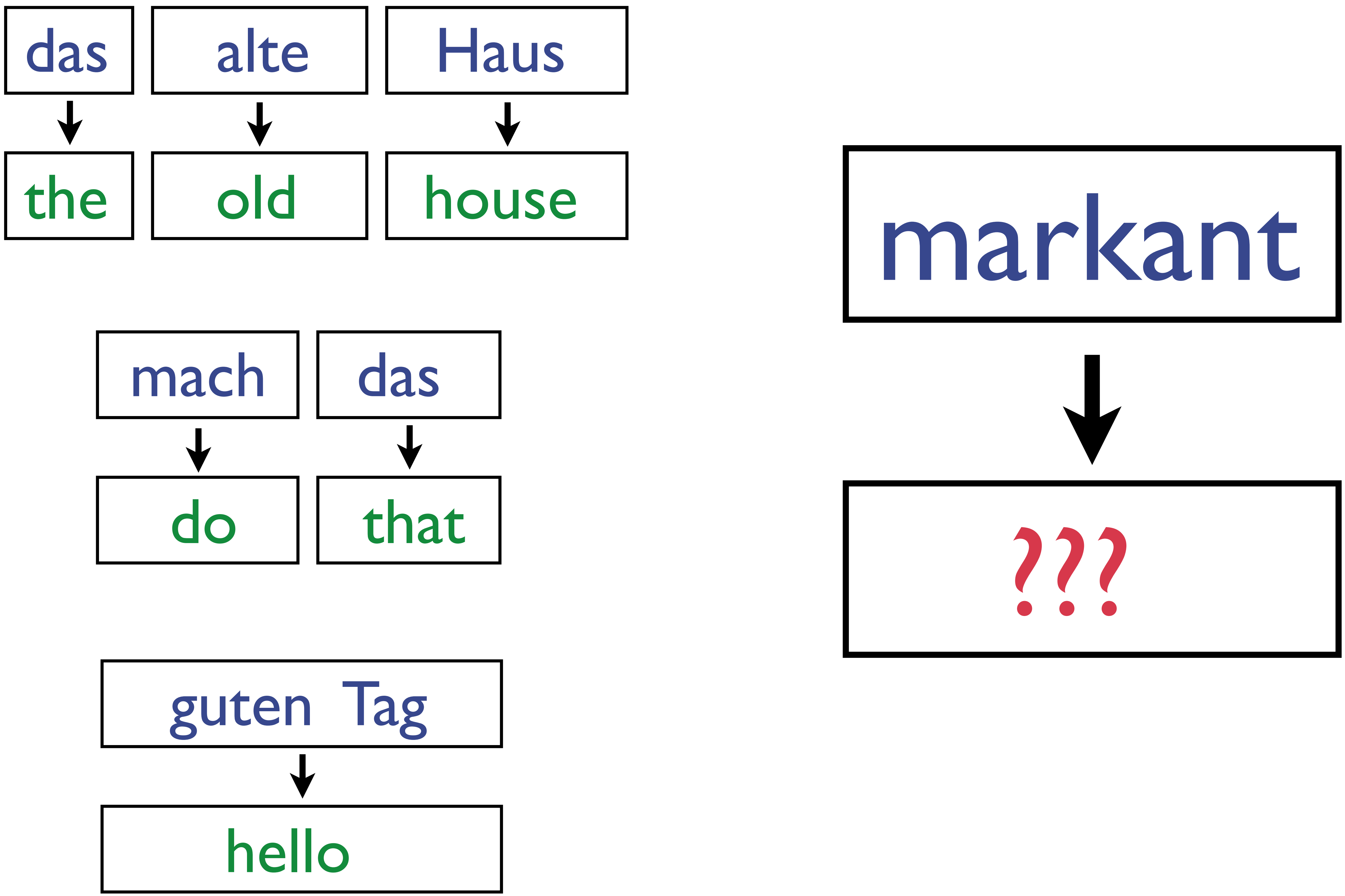

das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
das
that

das | alte | Haus
↓ | ↓ | ↓
the | old | house

mach | das
↓ | ↓
do | that

guten Tag
↓
hello

markant


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
markant
???

So far so good,

but....

das → the

alte → old

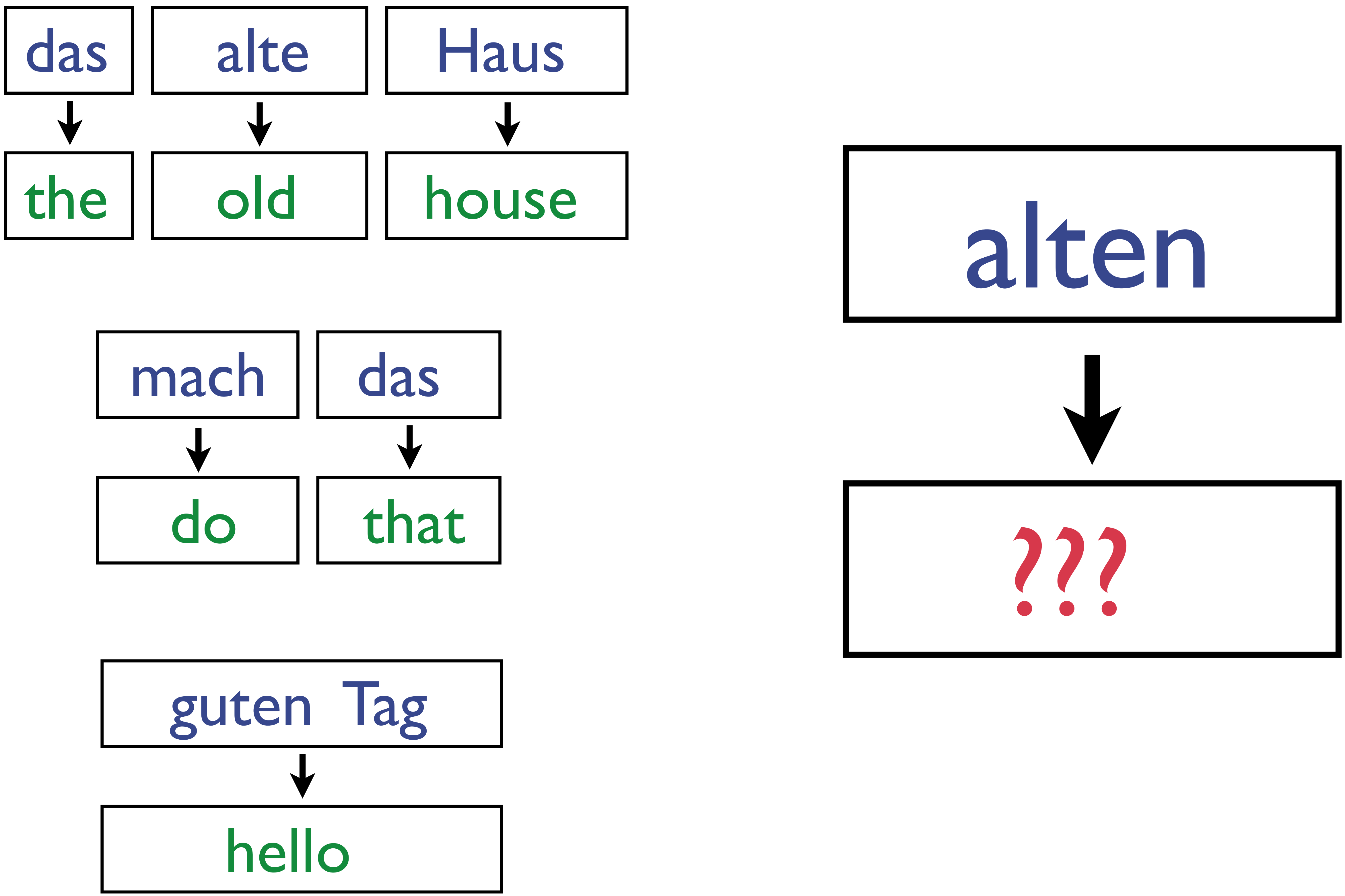
Haus → house

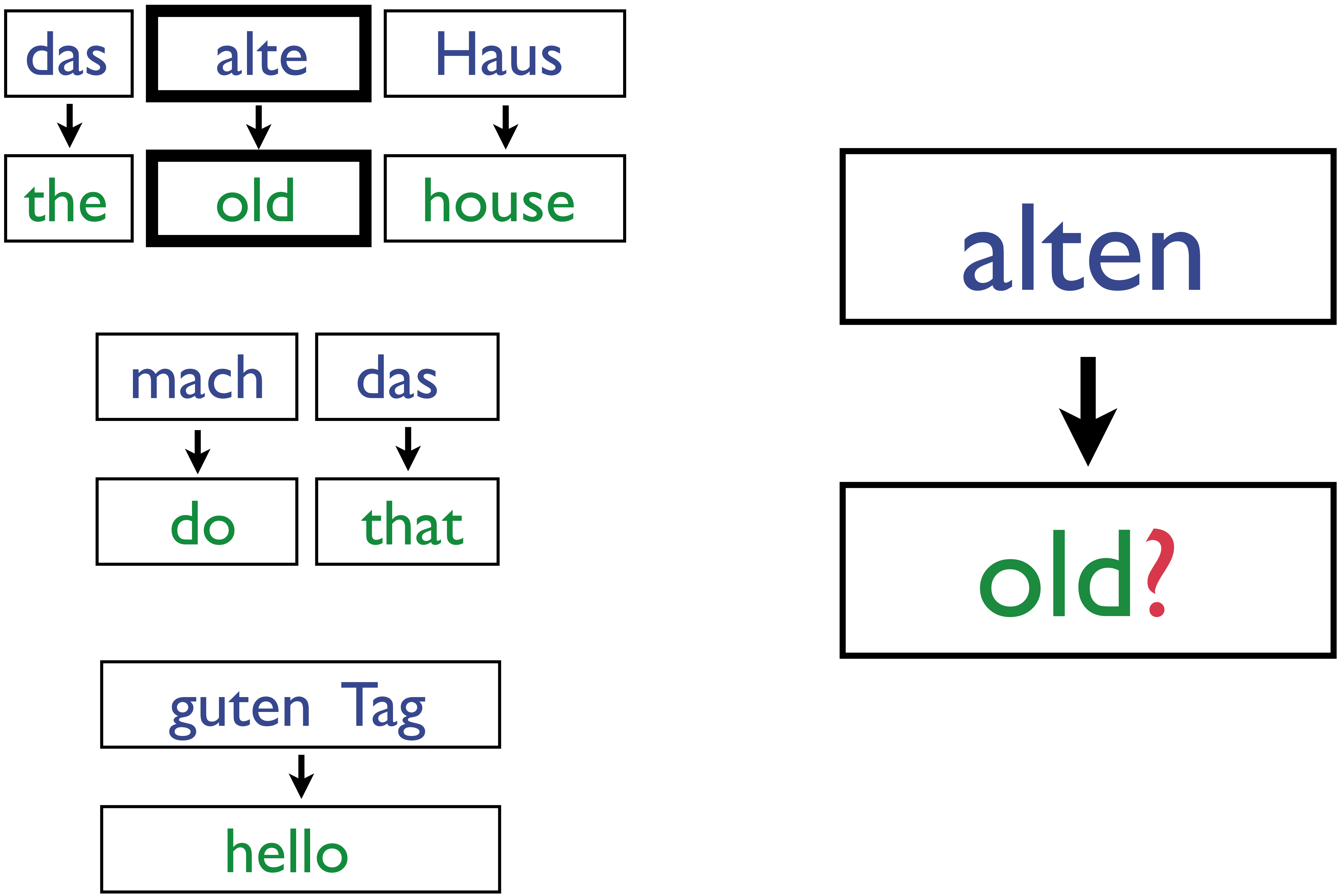
mach → do

das → that

guten Tag → hello

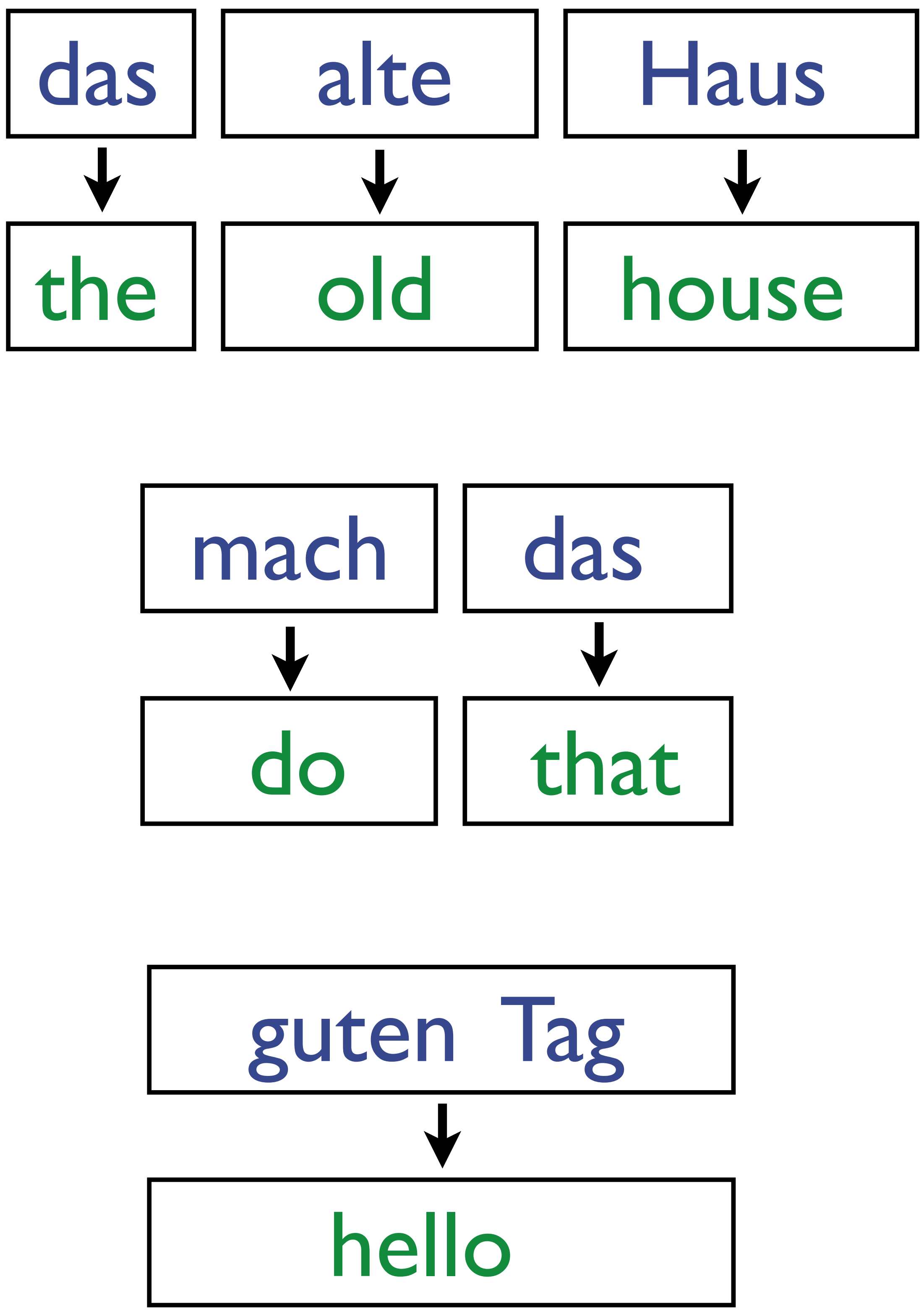
alten

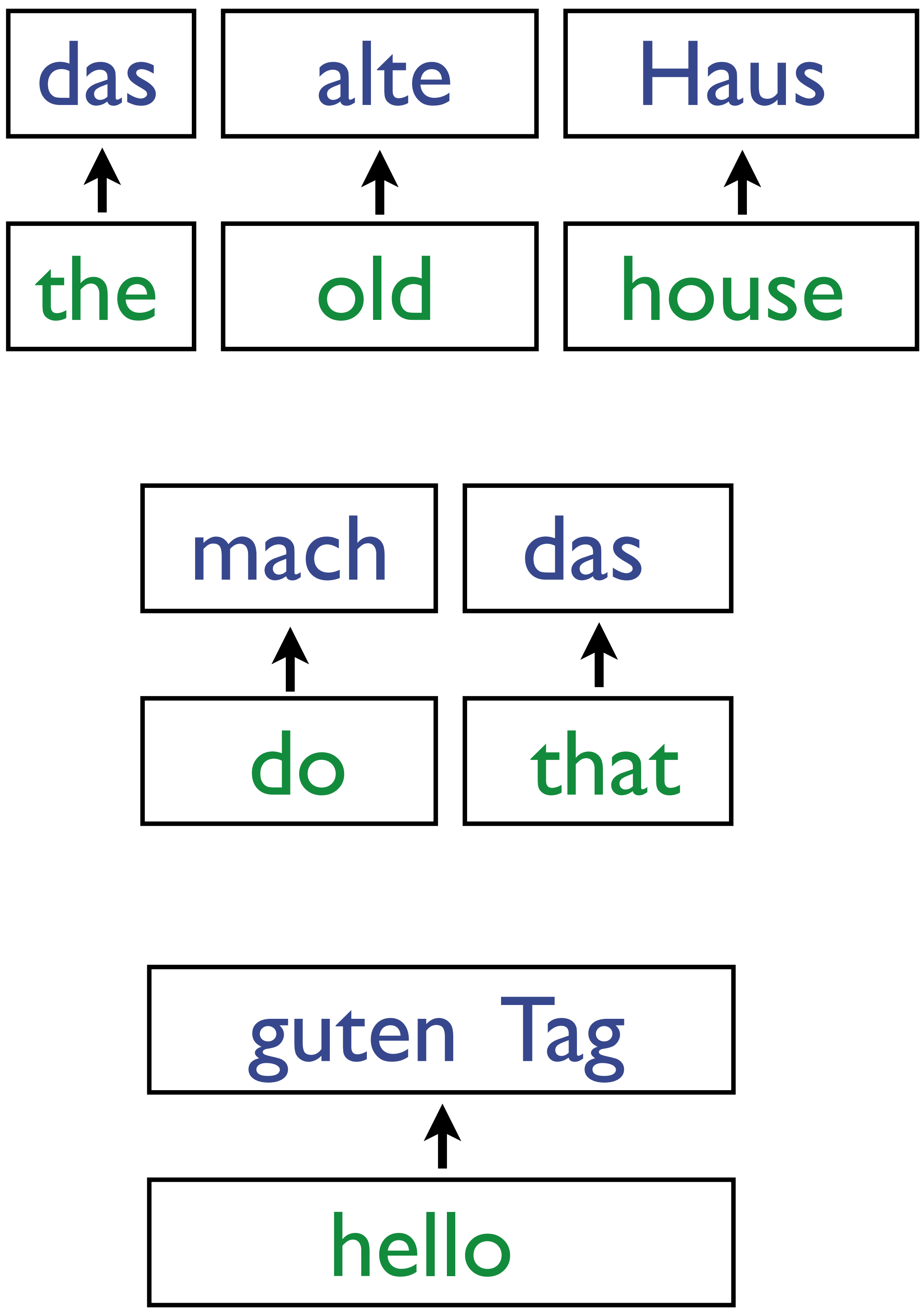

das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
alten
???

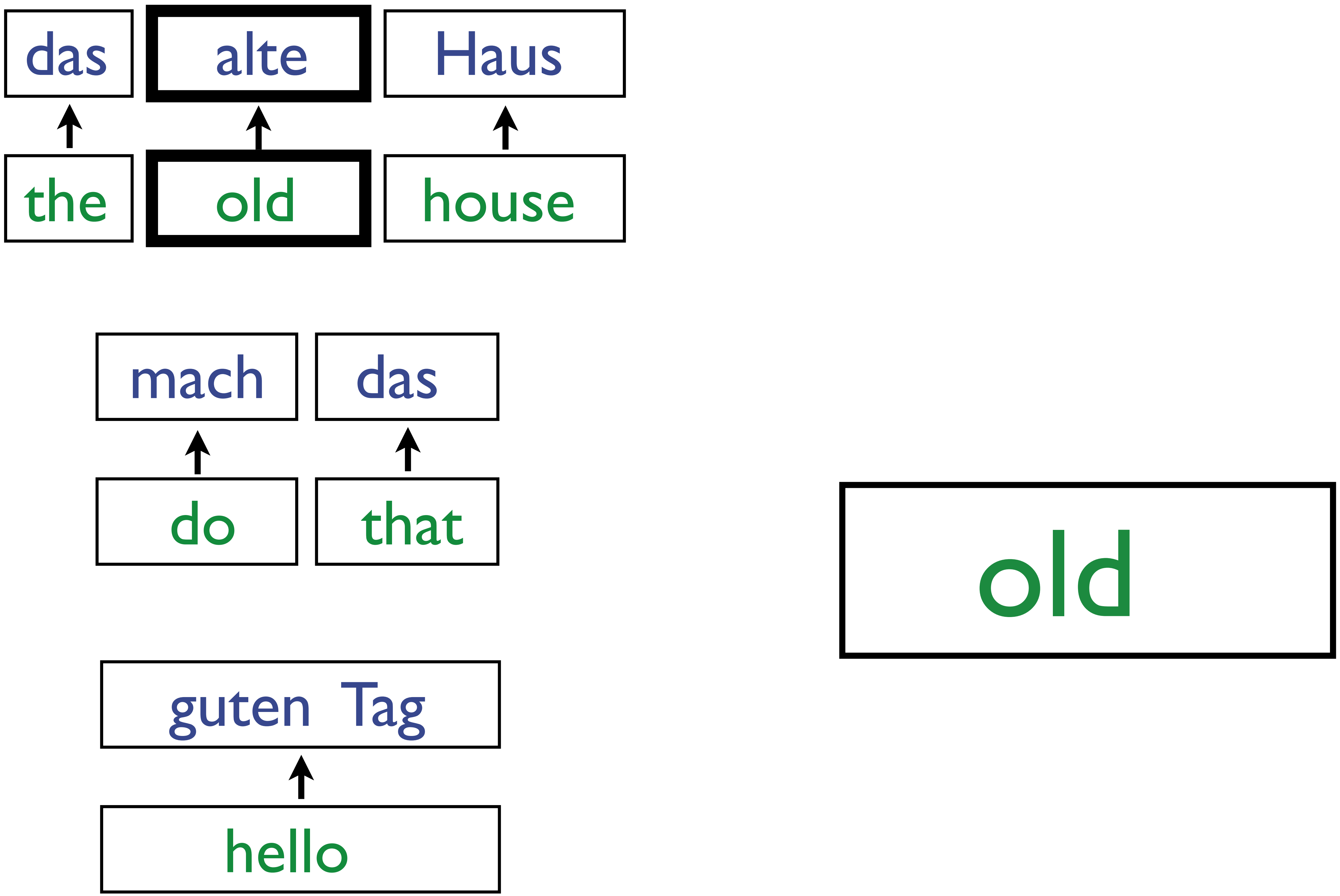

das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
alten
old?

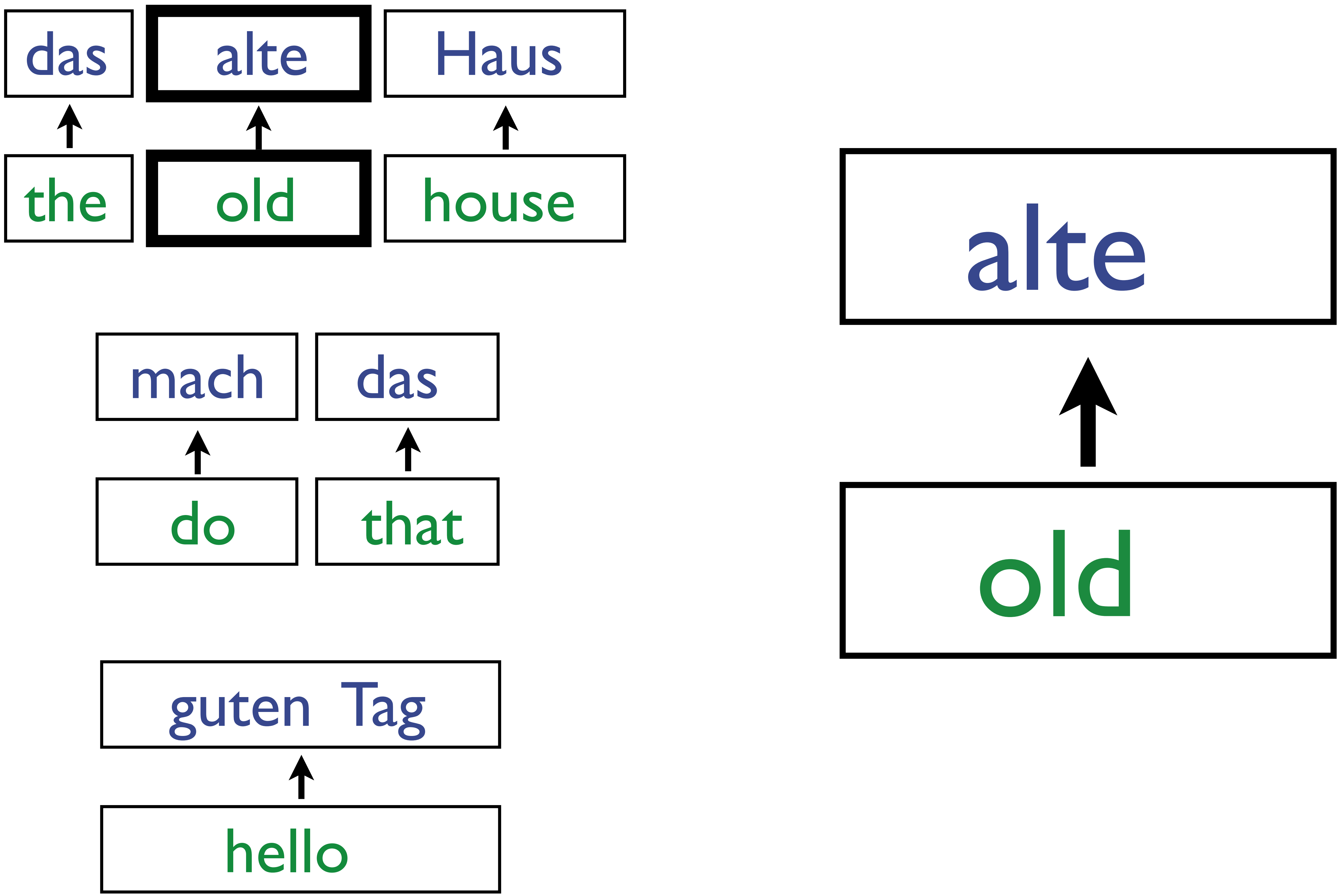
Problems

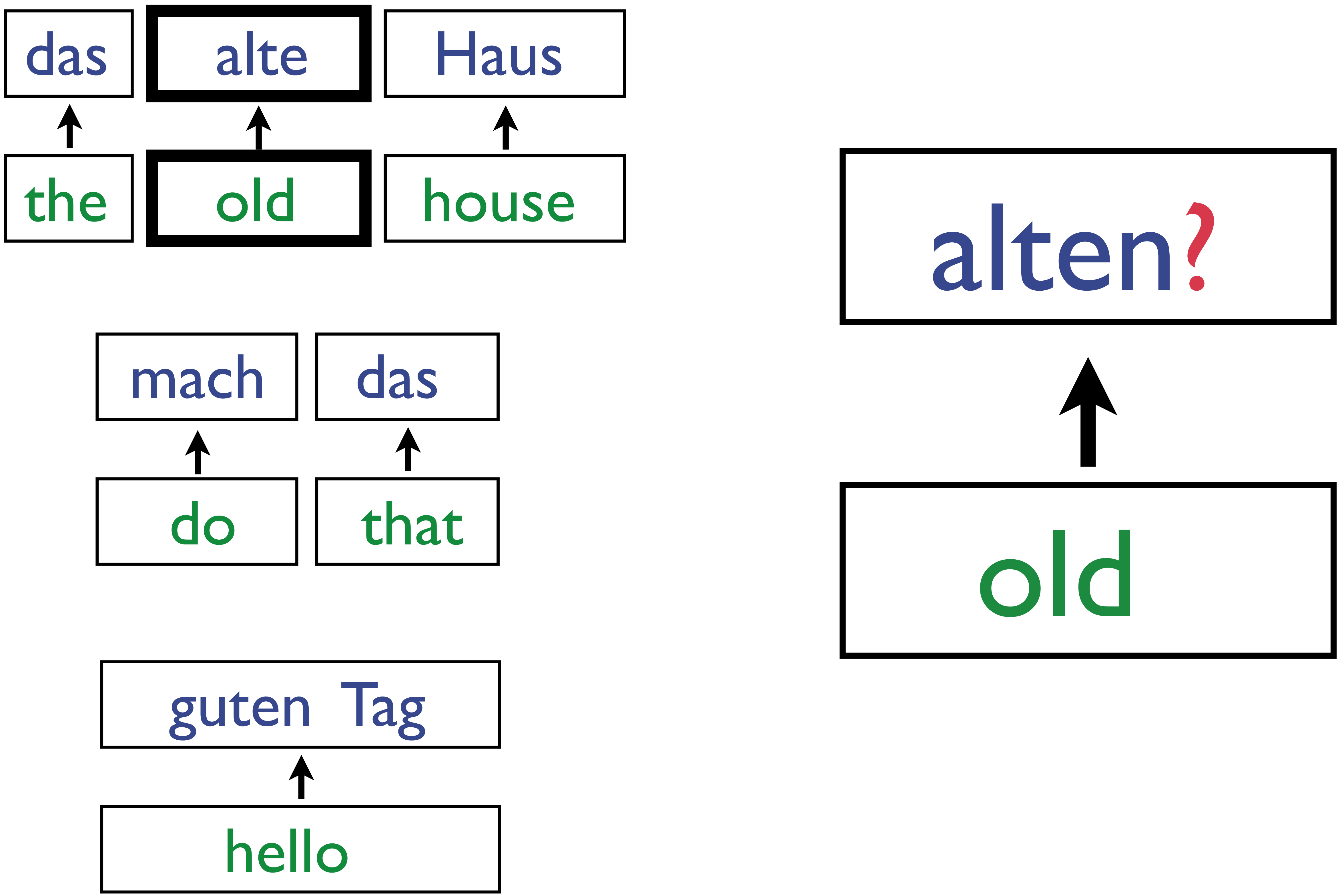
1. Source language inflectional richness.


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
old


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
alte
old


das
alte
Haus
the
old
house
mach
das
do
that
guten Tag
hello
alten?
old

Problems

1. Source language inflectional richness.

2. Target language inflectional richness.

Kopfschmerzen

↓

head ache

Bauchschmerzen

↓

abdominal pain

Rücken

↓

back

Kopf

↓

head

Rückenschmerzen

Kopfschmerzen

head ache

Bauchschmerzen

abdominal pain

Rücken

back

Kopf

head

Rückenschmerzen

???

Kopfschmerzen

↓

head ache

Bauch***schmerzen***

↓

abdominal ***pain***

Rücken

↓

back

Kopf

↓

head

Rückenschmerzen

↓

back pain

Kopf***schmerzen***

↓

head ***ache***

Bauchschmerzen

↓

abdominal pain

Rücken

↓

back

Kopf

↓

head

Rückenschmerzen

↓

back ache

Problems

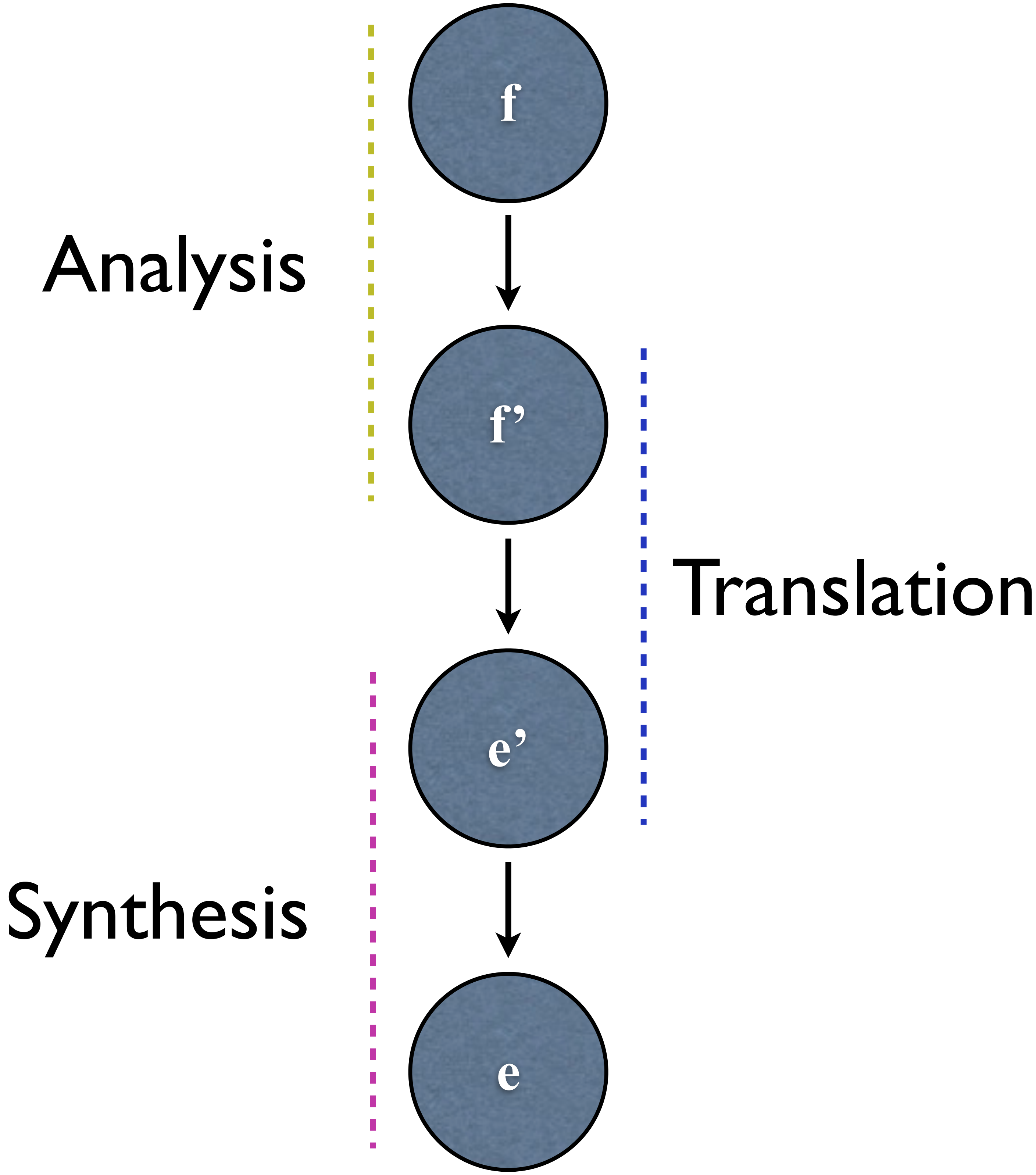
1. Source language inflectional richness.

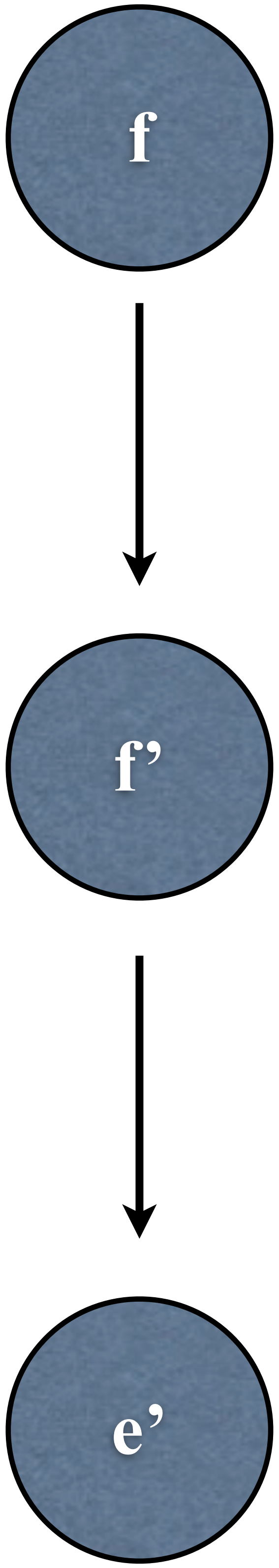
2. Target language inflectional richness.

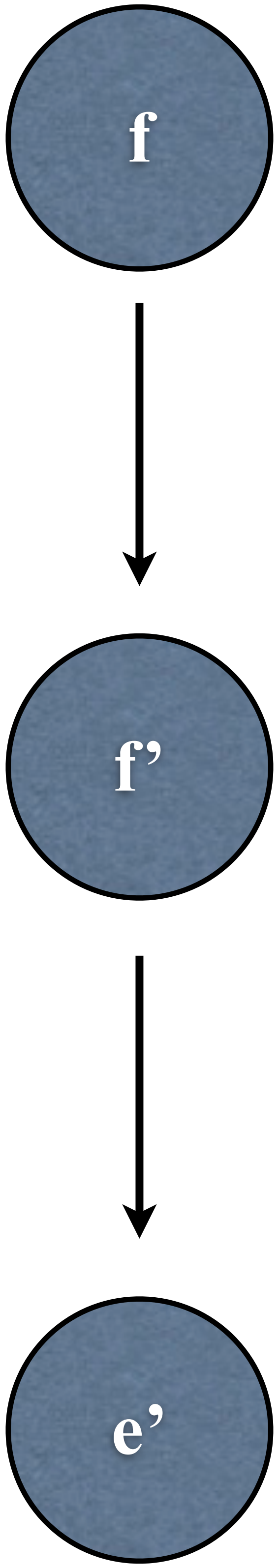
3. Source language sublexical semantic compositionality.

General Solution

MORPHOLOGY


f
Analysis
f'
Translation
e'
Synthesis
e


f
f’
e’



AlAbAmA

f

AlAbAmA

↓

f’

Al# Abama (looks like Al + OOV)

↓

e’

f AlAbAmA

f' Al# Abama (looks like Al + OOV)

e' the Ibama

But...Ambiguity!

- Morphology is an inherently ambiguous problem
  - Competing linguistic theories
  - Lexicalization
- Morphological analyzers (tools) make mistakes
- Are minimal linguistic morphemes the optimal morphemes for MT?

Problems

1. Source language inflectional richness.

2. Target language inflectional richness.

3. Source language sublexical semantic compositionality.

4. Ambiguity everywhere!

General Solution

MORPHOLOGY

+

PROBABILITY

Why probability?

- **Probabilistic models formalize uncertainty**
- e.g., words can be formed via a morphological derivation according to a joint distribution:

$$p(\text{word}, \text{derivation})$$

- The probability of a word is naturally defined as the marginal probability:

$$p(\text{word}) = \sum_{\text{derivation}} p(\text{word}, \text{derivation})$$

- Such a model can even be trained observing just words (EM!)

$$
\begin{aligned}
p(\text{derived}) = \\
& p(\text{derived}, \text{de+rive+d}) + \\
& p(\text{derived}, \text{derived+}\varnothing) + \\
& p(\text{derived}, \text{derive+d}) + \\
& p(\text{derived}, \text{deriv+ed}) + \ldots
\end{aligned}
$$

Outline

- Introduction: 4 problems
- Three probabilistic modeling solutions
    - Embracing uncertainty: multi-segmentations for decoding and learning
    - Rich morphology via sparse lexical features
    - Hierarchical Bayesian translation: infinite translation lexicons
- Conclusion

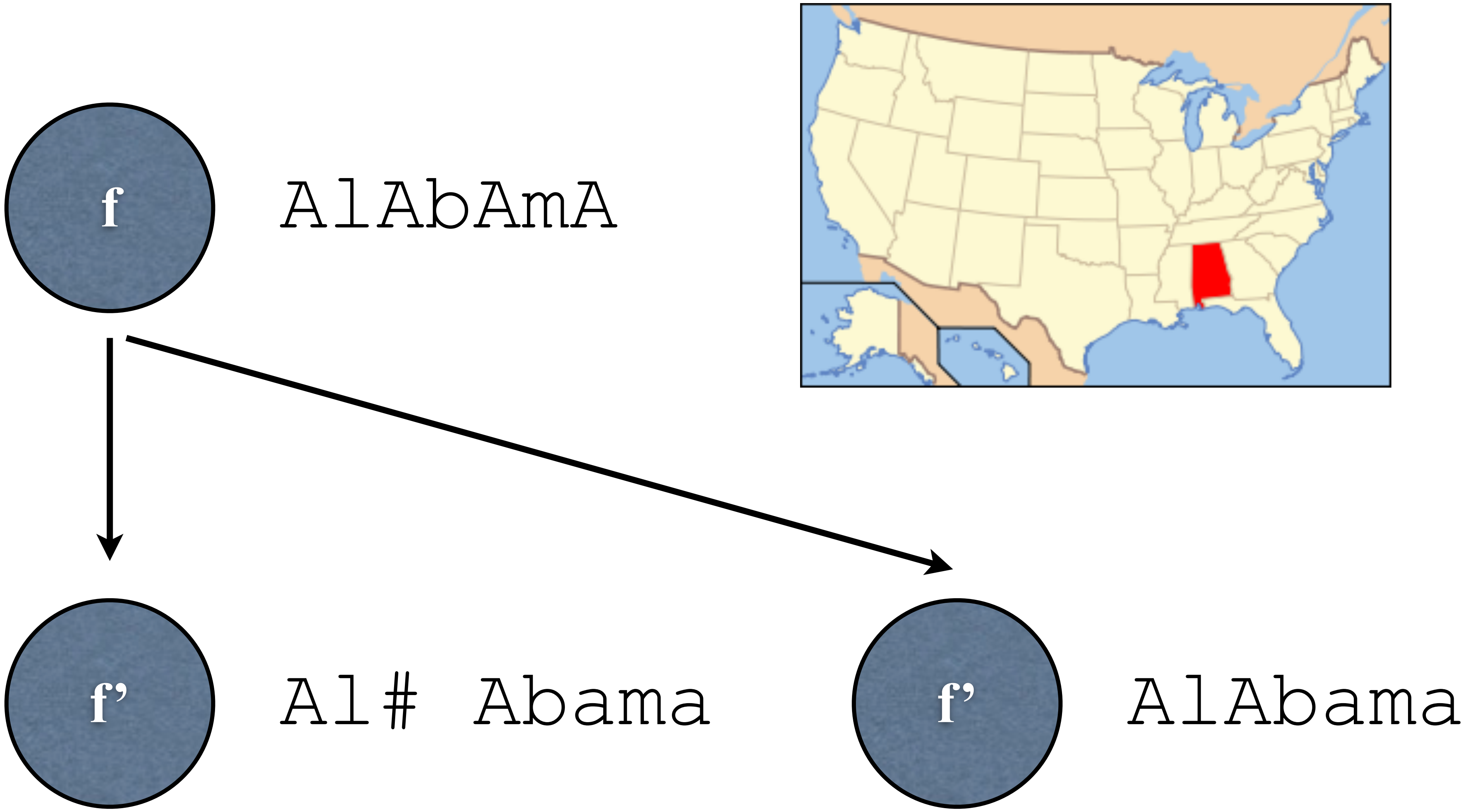
Outline

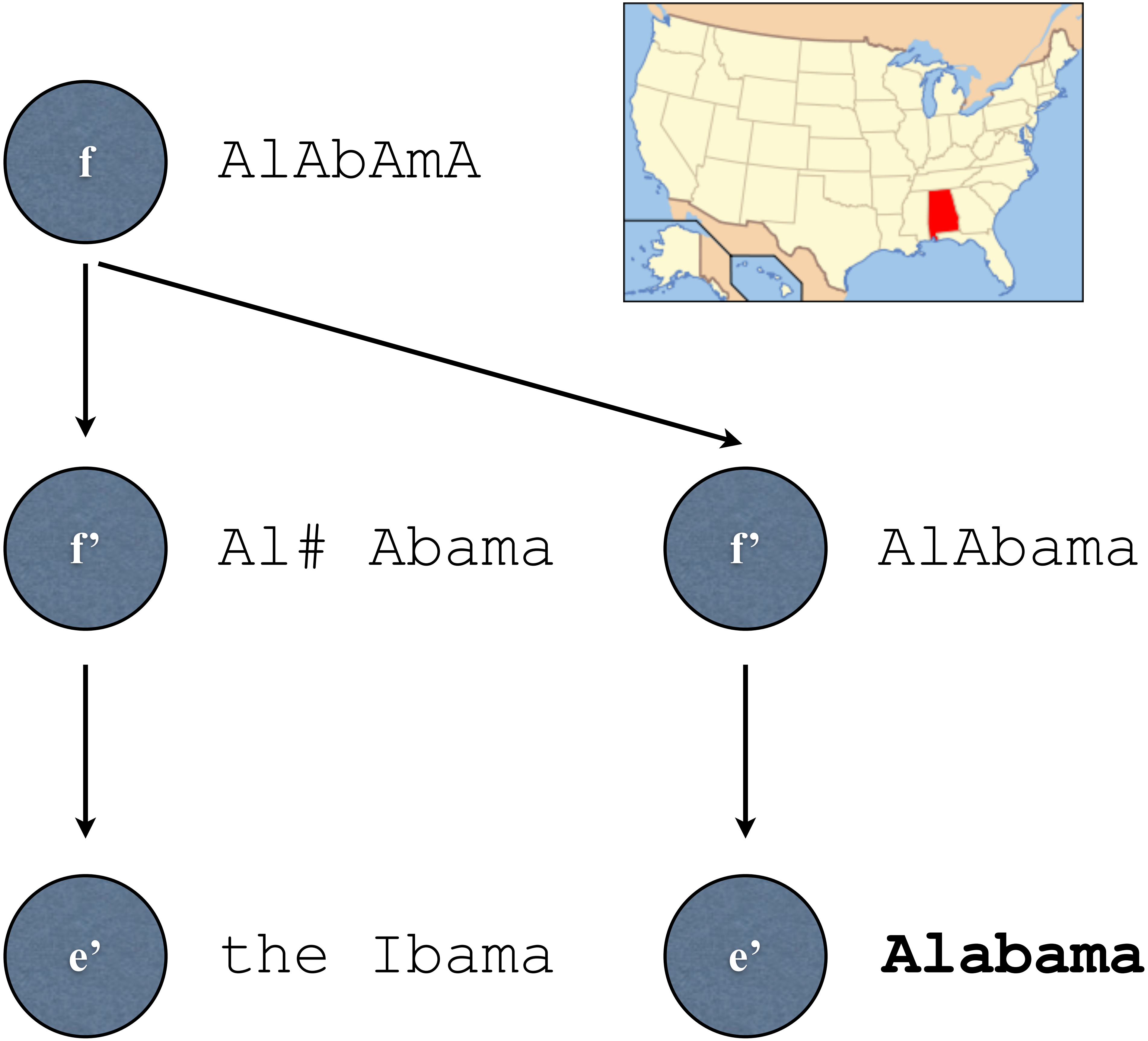
- Introduction: 4 problems
- Three probabilistic modeling solutions
  - Embracing uncertainty: multi-segmentations for decoding and learning
  - Rich morphology via sparse lexical features
  - Hierarchical Bayesian translation: infinite translation lexicons
- Conclusion


f

f

AlAbAmA


f
AlAbAmA
f’
Al# Abama
f’
AlAbama


f
AlAbAmA
f’
Al# Abama
f’
AlAbama
e’
the Ibama
e’
Alabama

Two problems

- We need to decode lots of similar source candidates efficiently
  - Lattice / confusion network decoding

    Kumar & Byrne (EMNLP, 2005), Bertoldi, Zens, Federico (ICAASP, 2007), Dyer et al. (ACL, 2008), *inter alia*

Two problems

- We need to decode lots of similar source candidates efficiently
    - Lattice / confusion network decoding

      Kumar & Byrne (EMNLP, 2005), Bertoldi, Zens, Federico (ICAASP, 2007), Dyer et al. (ACL, 2008), *inter alia*

- We need a model to generate a set of candidate sources
    - **What are the right candidates?**

Uncertainty is everywhere

Requirement: a probabilistic model $p(\mathbf{f}'|\mathbf{f})$ that transforms $\mathbf{f} \rightarrow \mathbf{f}'$

Possible solution: a discriminatively trained model, e.g., a CRF

Required data: example $(\mathbf{f},\mathbf{f}')$ pairs from a linguistic expert or other source

Uncertainty is everywhere

What is the best/right analysis ... for MT?

`AlAntxAbAt`

(DEF+election+PL)

Uncertainty is everywhere

What is the best/right analysis ... for MT?

`AlAntxAbAt`

(DEF+election+PL)

Some possibilities: Sadat & Habash (NAACL, 2007)

```
AlAntxAb +At
Al+ AntxAb +At
Al+ AntxAbAt
AlAntxAbAt
```

Uncertainty is everywhere

What is the best/right analysis ... for MT?

`AlAntxAbAt`

(DEF+election+PL)

Some possibilities: Sadat & Habash (NAACL, 2007)

```
AlAntxAb +At
Al+ AntxAb +At
Al+ AntxAbAt
AlAntxAbAt
```

Let's use them all!

Wait...multiple references?!?

- Train with EM variant
- Lattices can encode very large sets of references and support efficient inference

Dyer (NAACL, 2009), Dyer (thesis, 2010)

Wait...multiple references?!?

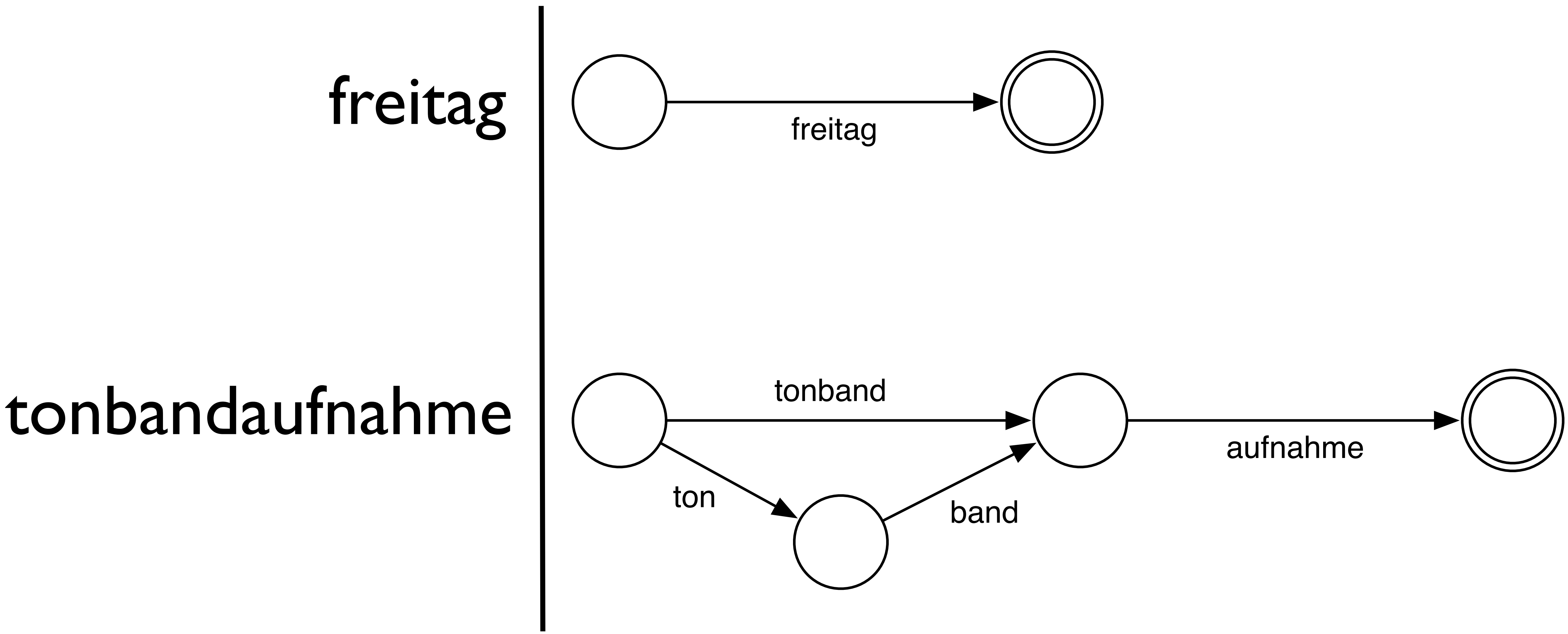
- Train with EM variant
- Lattices can encode very large sets of references and support efficient inference

Dyer (NAACL, 2009), Dyer (thesis, 2010)

- Bonus: annotation task is **much** simpler
  - Don’t know whether to label an example with A or B?
  - Label it with **both**!

Reference Segmentations

f **f’**



good phonotactics!

Rücken + schmerzen

Rückenschmerzen

Rückensc + hmerzen

Rü + cke + nschme + rzen

bad phonotactics!

Phonotactic features!

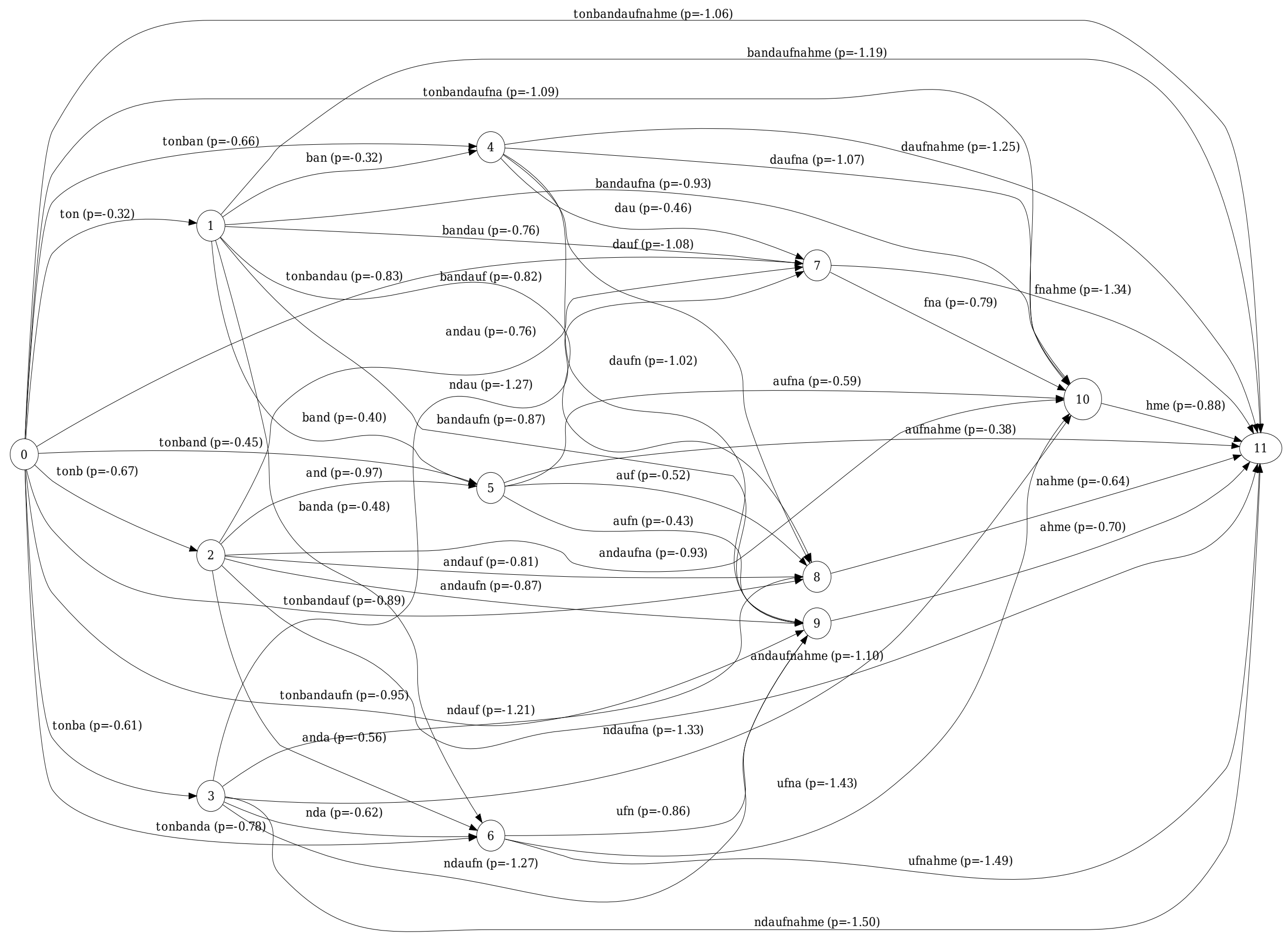
Just 20 features

- Phonotactic probability
- Lexical features (in vocab, OOV)
- Lexical frequencies
- Is high frequency?
- Segment length
- ...

https://github.com/redpony/cdec/tree/master/compound-split

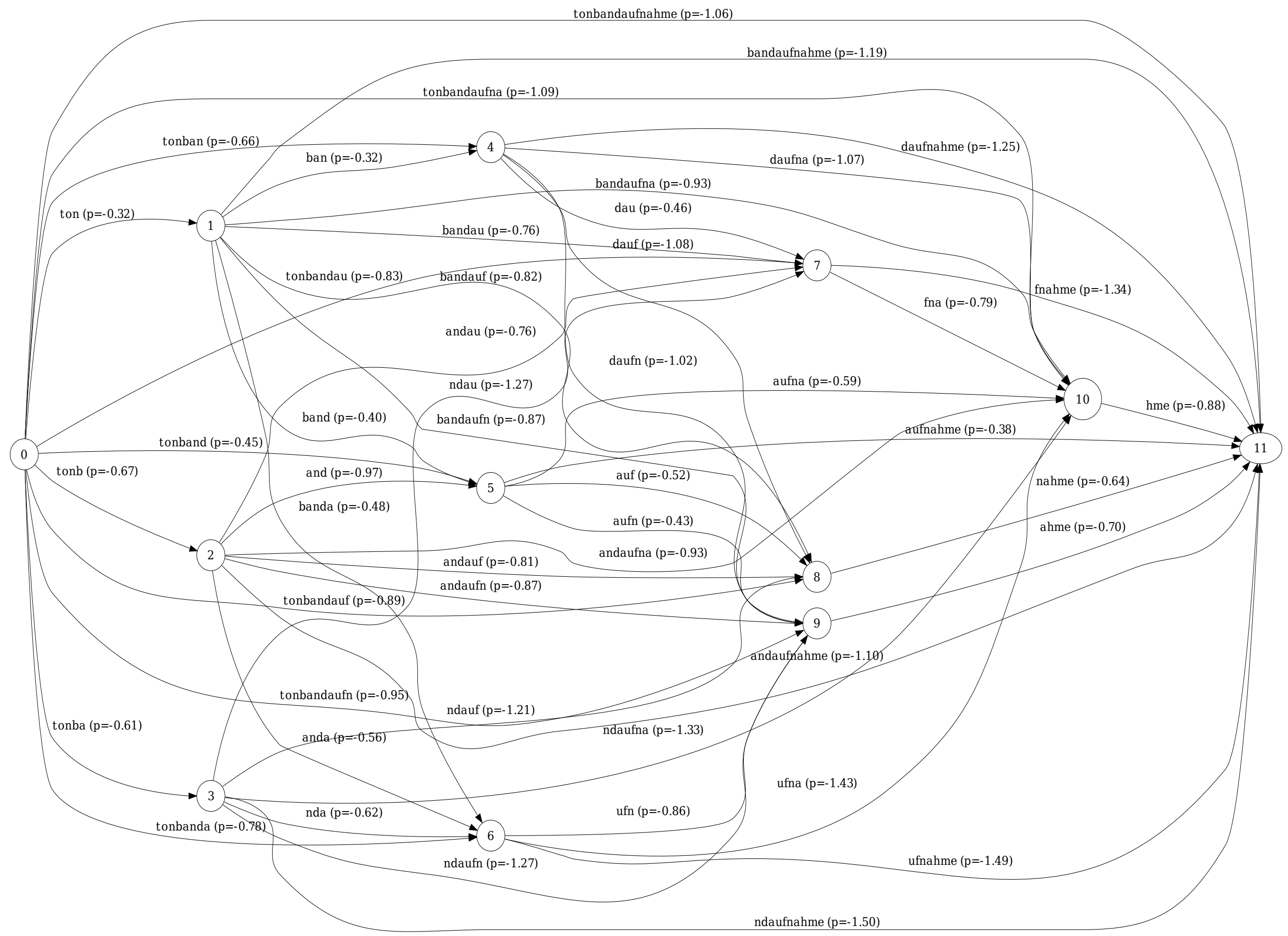
Input: **tonbandaufnahme**

Input: **tonbandaufnahme**

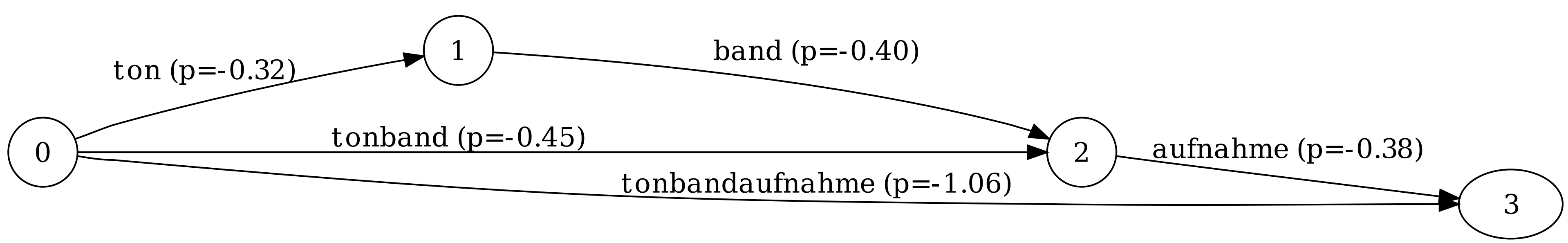


Input: **tonbandaufnahme**

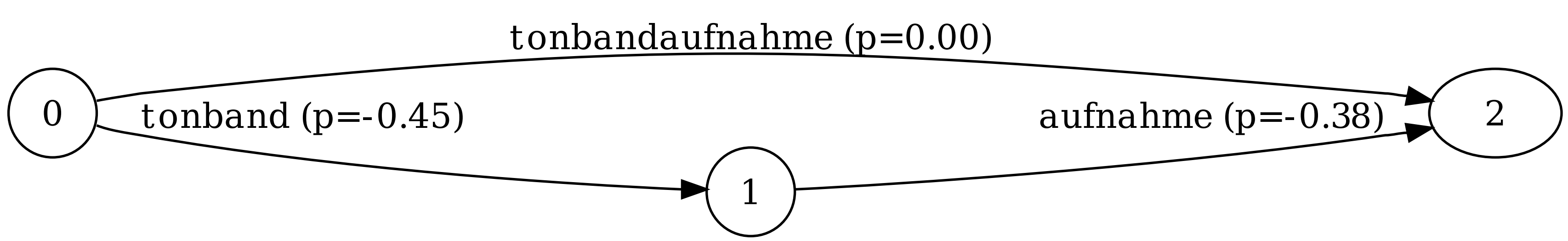
a=∞

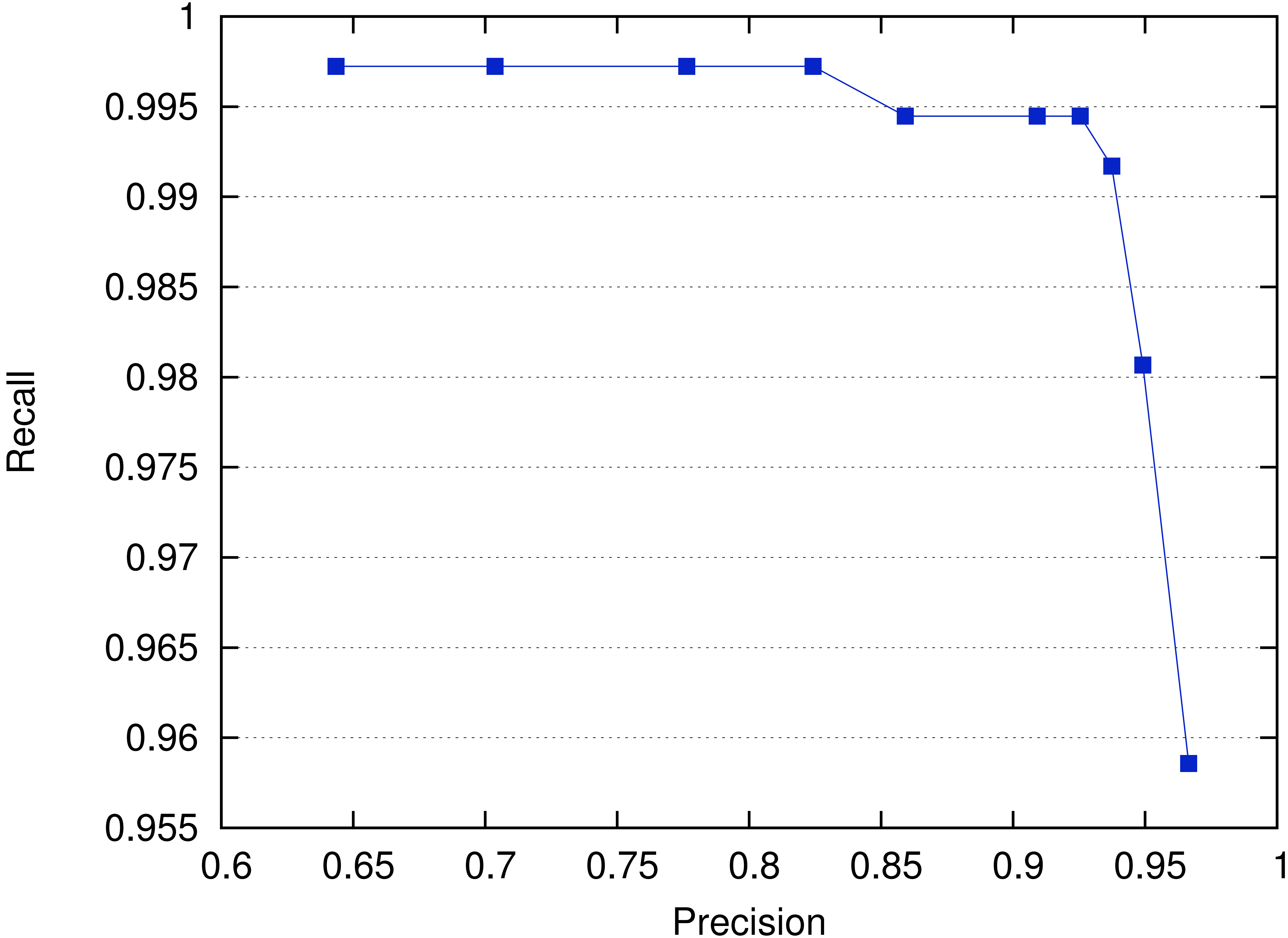


a=0.4



a=0.2




1
0.995
0.99
0.985
0.98
0.975
0.97
0.965
0.96
0.955
Recall
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Precision

Translation Evaluation

Input	BLEU	TER
Unsegmented	20.8	61.0
1-best segmentation	20.3	60.2
Lattice (a=0.2)	**21.5**	**59.8**

in police raids found illegal guns , ammunition ***stahlkern*** , ***laserzielfernrohr*** and a machine gun .

in police raids found with illegal guns and ammunition ***steel core*** , a ***laser objective telescope*** and a machine gun .

REF:

police raids found illegal guns , ***steel core*** ammunition , a ***laser scope*** and a machine gun .

Outline

- Introduction: 4 problems
- Three probabilistic modeling solutions
  - Embracing uncertainty: multi-segmentations for decoding and learning
  - Rich morphology via sparse lexical features
  - Hierarchical Bayesian translation: infinite translation lexicons
- Conclusion

What do we see when we look inside the IBM models?

(or any multinomial-based generative model...like parsing models!)

What do we see when we look inside the IBM models?

(or any multinomial-based generative model...like parsing models!)

old

altes	0.3
alte	0.1
alt	0.2
alter	0.1
gammelig	0.1
gammeliges	0.1

car

Wagen	0.2
Auto	0.6
PKW	0.2

What do we see when we look inside the IBM models?

(or any multinomial-based generative model...like parsing models!)

old

altes	0.3
alte	0.1
alt	0.2
alter	0.1
gammelig	0.1
gammeliges	0.1

car

Wagen	0.2
Auto	0.6
PKW	0.2

DLVM for Translation

Addresses problems:

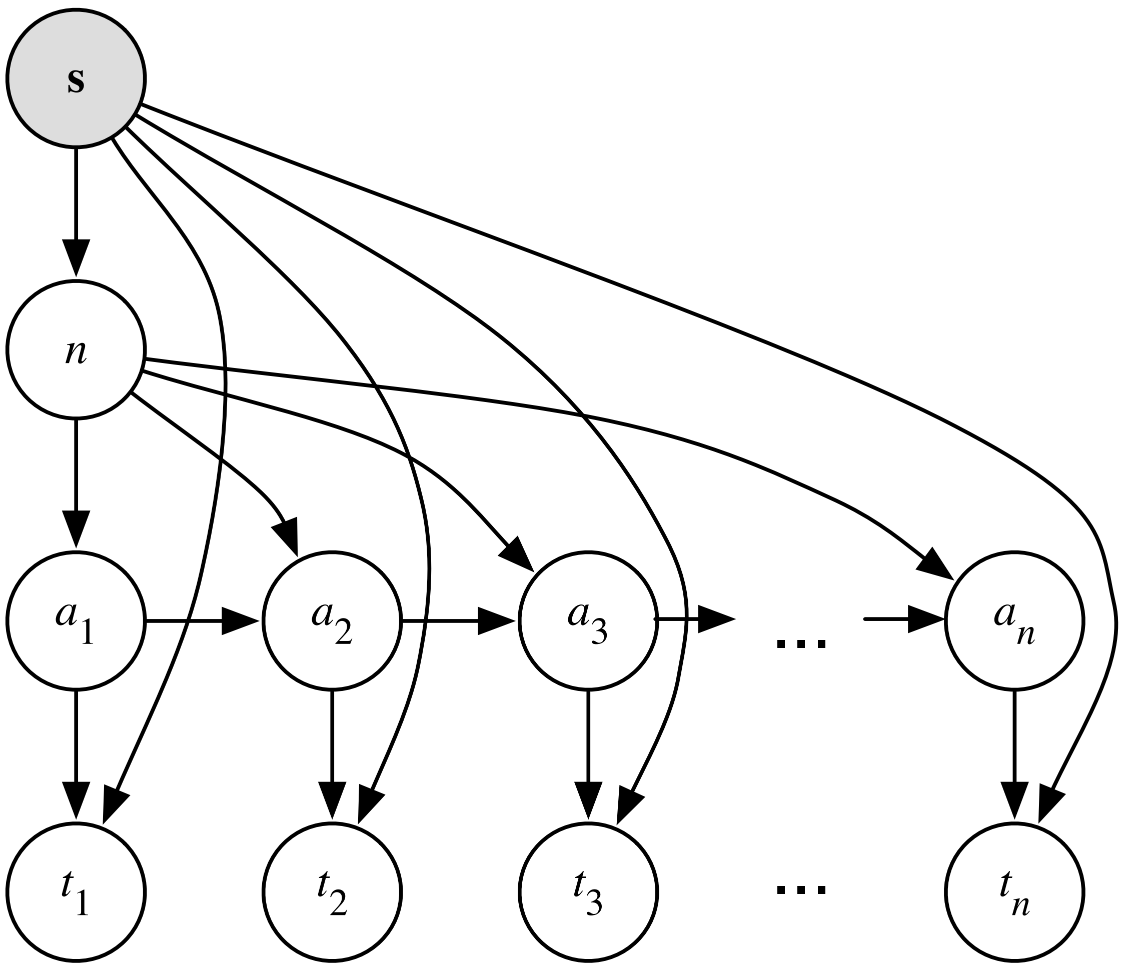
1. Source language inflectional richness.
2. Target language inflectional richness.

How?

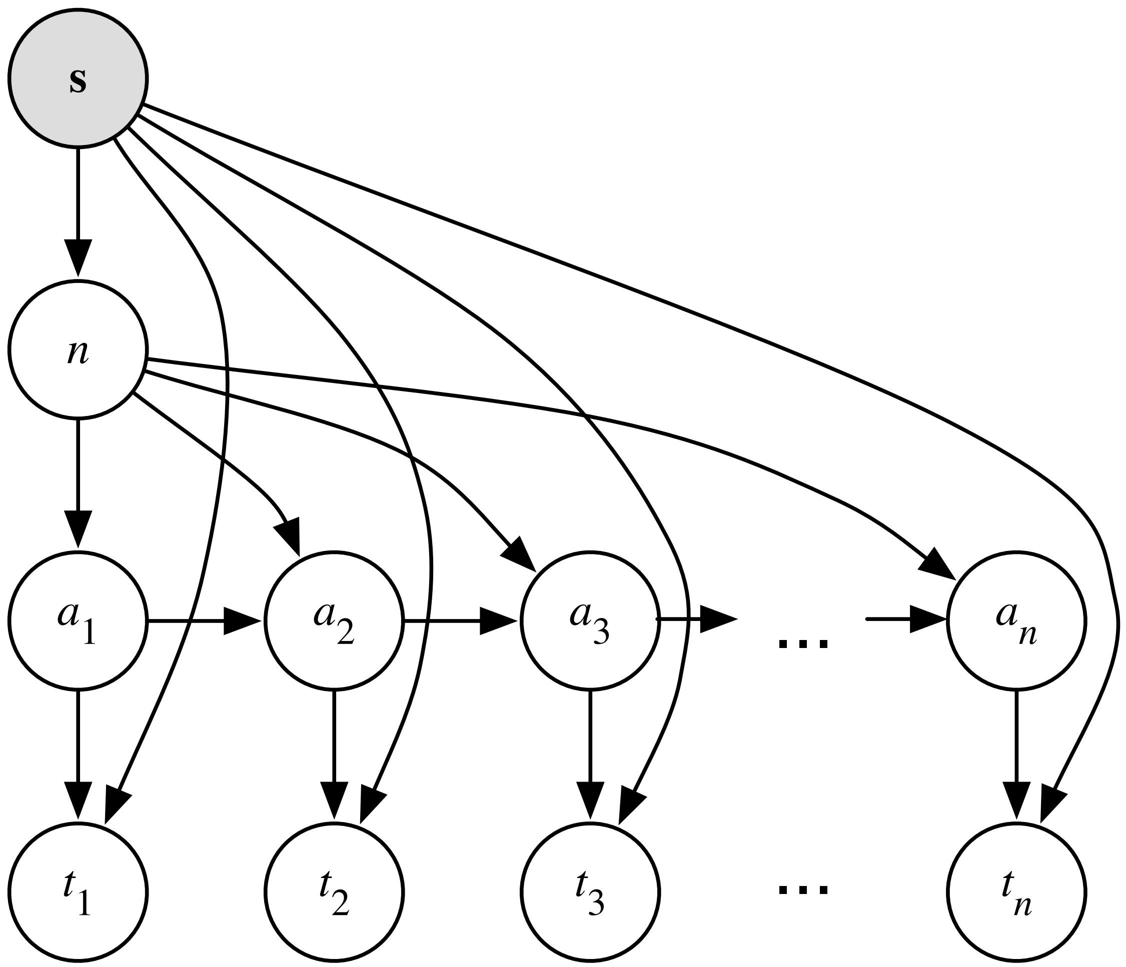
1. Replace the locally normalized multinomial parameterization in a translation model $p(\mathbf{e} \mid \mathbf{f})$ with a globally normalized log-linear model.

2. Add lexical association features sensitive to sublexical units.

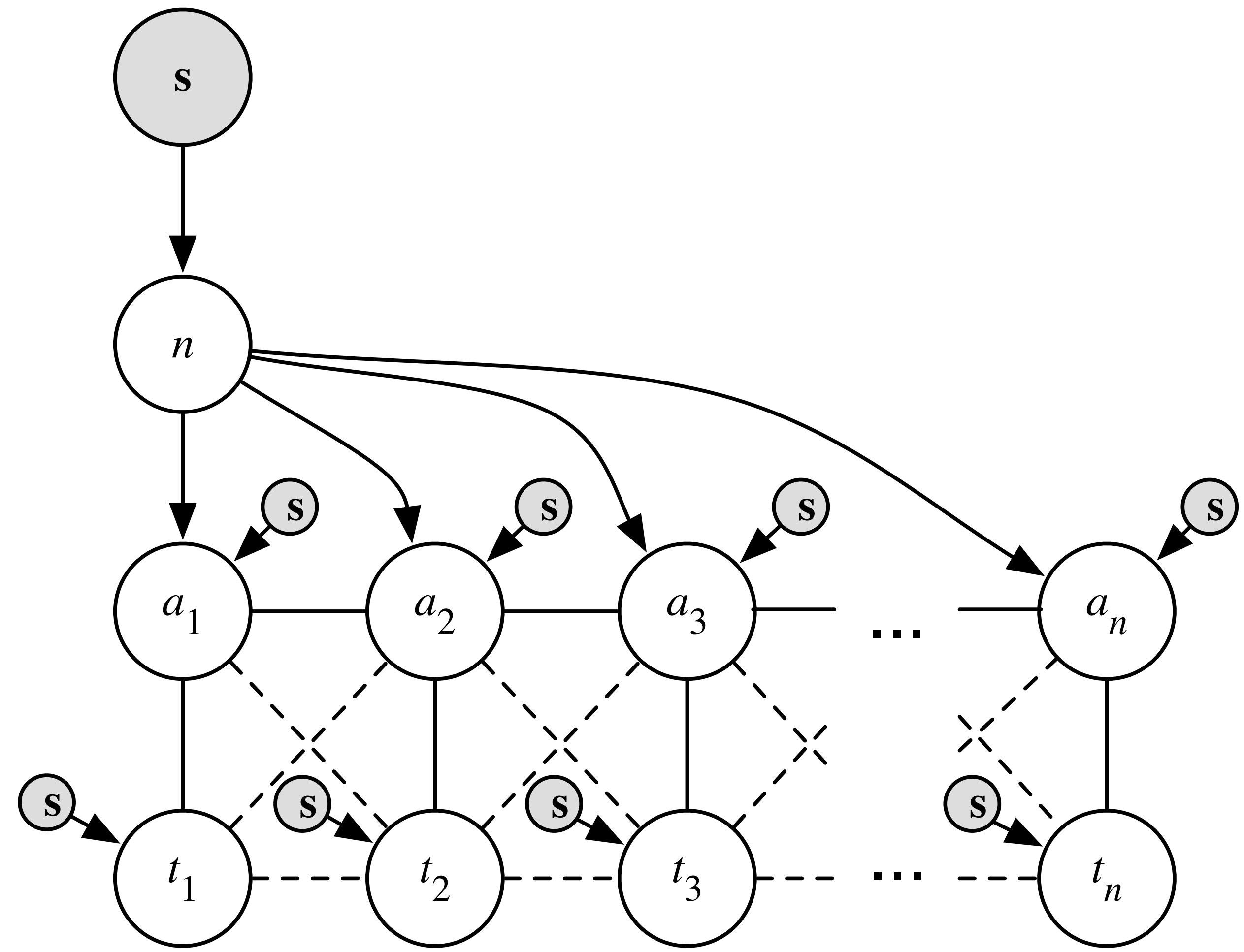
C. Dyer, J. Clark, A. Lavie, and N. Smith (in review)



Fully directed model (Brown et al., 1993; Vogel et al., 1996; Berg-Kirkpatrick et al., 2010)



Fully directed model (Brown et al., 1993;
Vogel et al., 1996; Berg-Kirkpatrick et al., 2010)



Our model

old		
	altes	0.3
	alte	0.1
	alt	0.2
	alter	0.1
	gammelig	0.1
	gammeliges	0.1

car		
	Wagen	0.2
	Auto	0.6
	PKW	0.2

old		
	altes	0.3
	alte	0.1
	alt	0.2
	alter	0.1
	gammelig	0.1
	gammeliges	0.1

car		
	Wagen	0.2
	Auto	0.6
	PKW	0.2

New model:

$$score(\mathbf{e},\mathbf{f}) = 0.2h_1(\mathbf{e},\mathbf{f}) + 0.9h_2(\mathbf{e},\mathbf{f}) + 1.3h_1(\mathbf{e},\mathbf{f}) + \ldots$$

old

alt+ $\Omega^{[0,2]}$

gammelig+ $\Omega^{[0,2]}$

old

altes	0.3
alte	0.1
alt	0.2
alter	0.1
gammelig	0.1
gammeliges	0.1

car

Wagen	0.2
Auto	0.6
PKW	0.2

New model:

$$score(\mathbf{e},\mathbf{f}) = 0.2h_1(\mathbf{e},\mathbf{f}) + 0.9h_2(\mathbf{e},\mathbf{f}) + 1.3h_1(\mathbf{e},\mathbf{f}) + \ldots$$

old

alt+ $\Omega^{[0,2]}$

gammelig+$\Omega^{[0,2]}$

(~ Incremental vs. realizational)

Sublexical Features

každoroční → **annual**

ID$_{\text{každoroční_annual}}$

PREFIX$_{\text{kaž_ann}}$

PREFIX$_{\text{každ_annu}}$

PREFIX$_{\text{každo_annua}}$

SUFFIX$_{\text{í_l}}$

SUFFIX$_{\text{ní_al}}$

Sublexical Features

každoroční → **annually**

ID$_{\text{každoroční_annually}}$

PREFIX$_{\text{kaž_ann}}$

PREFIX$_{\text{každ_annu}}$

PREFIX$_{\text{každo_annua}}$

SUFFIX$_{\text{í_y}}$

SUFFIX$_{\text{ní_ly}}$

Sublexical Features

každoročního → **annually**

ID$_{\text{každoročního_annually}}$

PREFIX$_{\text{kaž_ann}}$

PREFIX$_{\text{každ_annu}}$

PREFIX$_{\text{každo_annua}}$

SUFFIX$_{\text{o_y}}$

SUFFIX$_{\text{ho_ly}}$

Sublexical Features

každoročního → **annually**

ID$_{\text{každoročního_annually}}$

PREFIX$_{\text{kaž_ann}}$

PREFIX$_{\text{každ_annu}}$

PREFIX$_{\text{každo_annua}}$

Abstract away from inflectional variation!

SUFFIX$_{\text{o_y}}$

SUFFIX$_{\text{ho_ly}}$

Evaluation

- Given a parallel corpus (no supervised alignments!), we can infer
  - The weights in the log-linear translation model
  - The MAP alignment
  - The model is a translation model, but we evaluate it as applied to **alignment**

Alignment Evaluation

		AER
Model 4	e|f	24.8
	f|e	33.6
	sym.	23.4
DLVM	e|f	21.9
	f|e	29.3
	sym.	**20.5**

Czech-English, 3.1M words training, 525 sentences gold alignments.

Translation Evaluation

Alignment	BLEU ↑	METEOR ↑	TER ↓
Model 4	$16.3_{\sigma=0.2}$	$46.1_{\sigma=0.1}$	$67.4_{\sigma=0.3}$
Our model	$16.5_{\sigma=0.1}$	$46.8_{\sigma=0.1}$	$67.0_{\sigma=0.2}$
Both	$\mathbf{17.4}_{\sigma=0.1}$	$\mathbf{47.7}_{\sigma=0.1}$	$\mathbf{66.3}_{\sigma=0.5}$

Czech-English, WMT 2010 test set, 1 reference

Outline

- Introduction: 4 problems
- Three probabilistic modeling solutions
  - Embracing uncertainty: multi-segmentations for decoding and learning
  - Rich morphology via sparse lexical features
  - Hierarchical Bayesian translation: infinite translation lexicons
- Conclusion

Bayesian Translation

Addresses problems:

2. Target language inflectional richness.

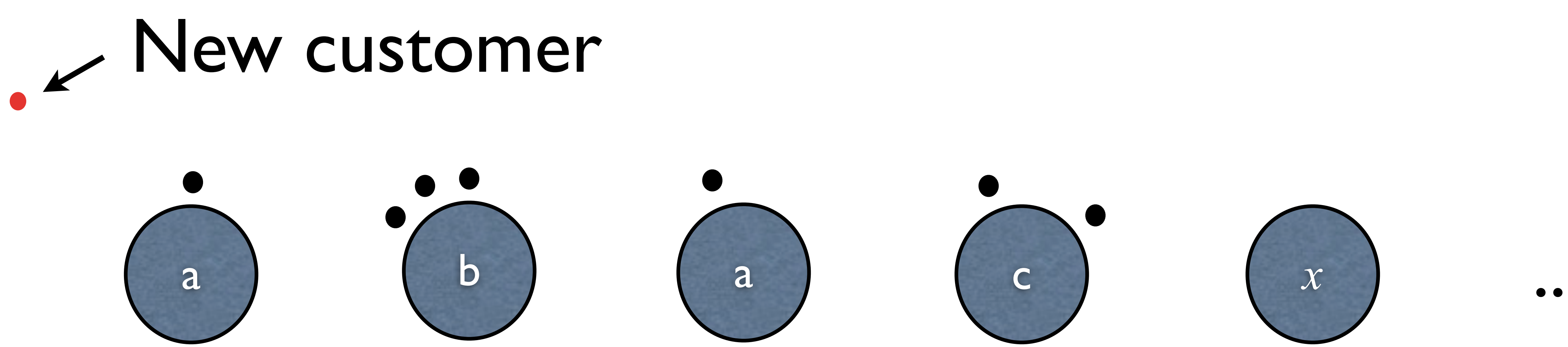
How?

1. Replace multinomials in a lexical translation model with a process that generates target language lexical items by combining stems and suffixes.

2. Fully inflected forms can be generated, but a hierarchical prior backs off to a component-wise generation.

Chinese Restaurant Process

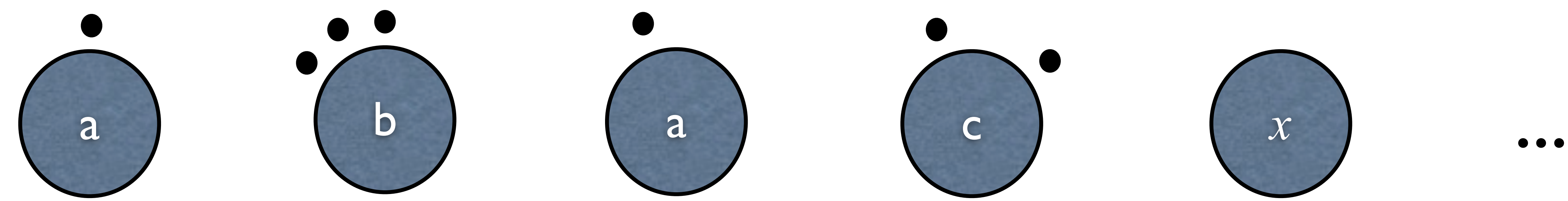
Chinese Restaurant Process



Chinese Restaurant Process

$$\frac{1}{7+\alpha} \qquad \frac{3}{7+\alpha} \qquad \frac{1}{7+\alpha} \qquad \frac{2}{7+\alpha} \qquad \frac{\alpha P_0(x)}{7+\alpha}$$

Chinese Restaurant Process



$$\frac{1}{7+\alpha} \qquad \frac{3}{7+\alpha} \qquad \frac{1}{7+\alpha} \qquad \frac{2}{7+\alpha} \qquad \frac{\alpha P_0(x)}{7+\alpha}$$

α "Concentration" parameter

$P_0(x)$ Base distribution

old

altes	0.3
alte	0.1
alt	0.2
alter	0.1
gammelig	0.1
gammeliges	0.1

car

Wagen	0.2
Auto	0.6
PKW	0.2

old

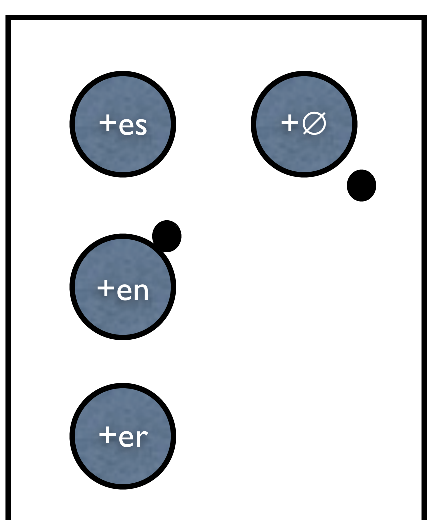
altes	0.3
alte	0.1
alt	0.2
alter	0.1
gammelig	0.1
gammeliges	0.1

car

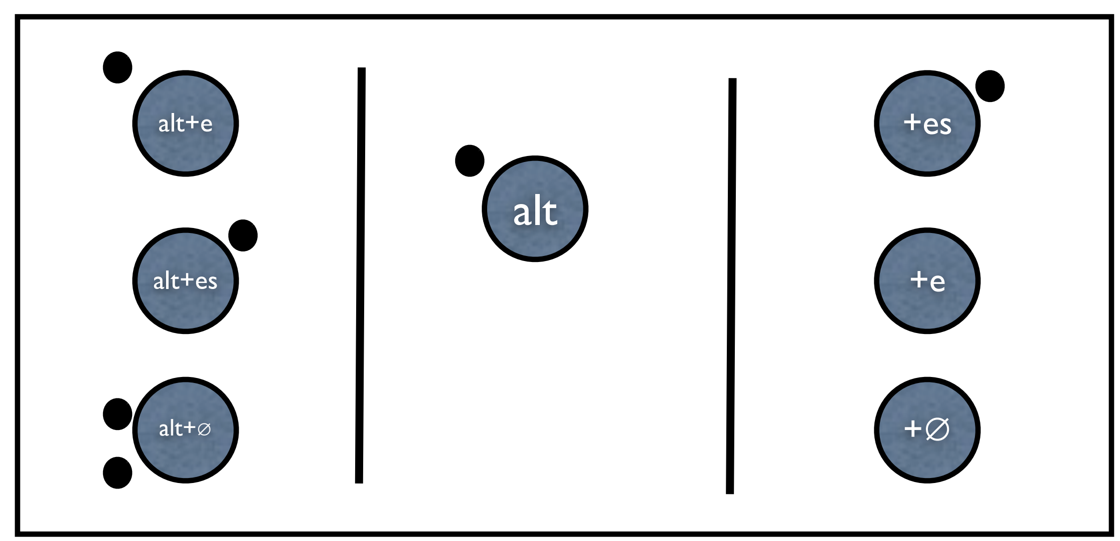
Wagen	0.2
Auto	0.6
PKW	0.2

New model:

suffixes

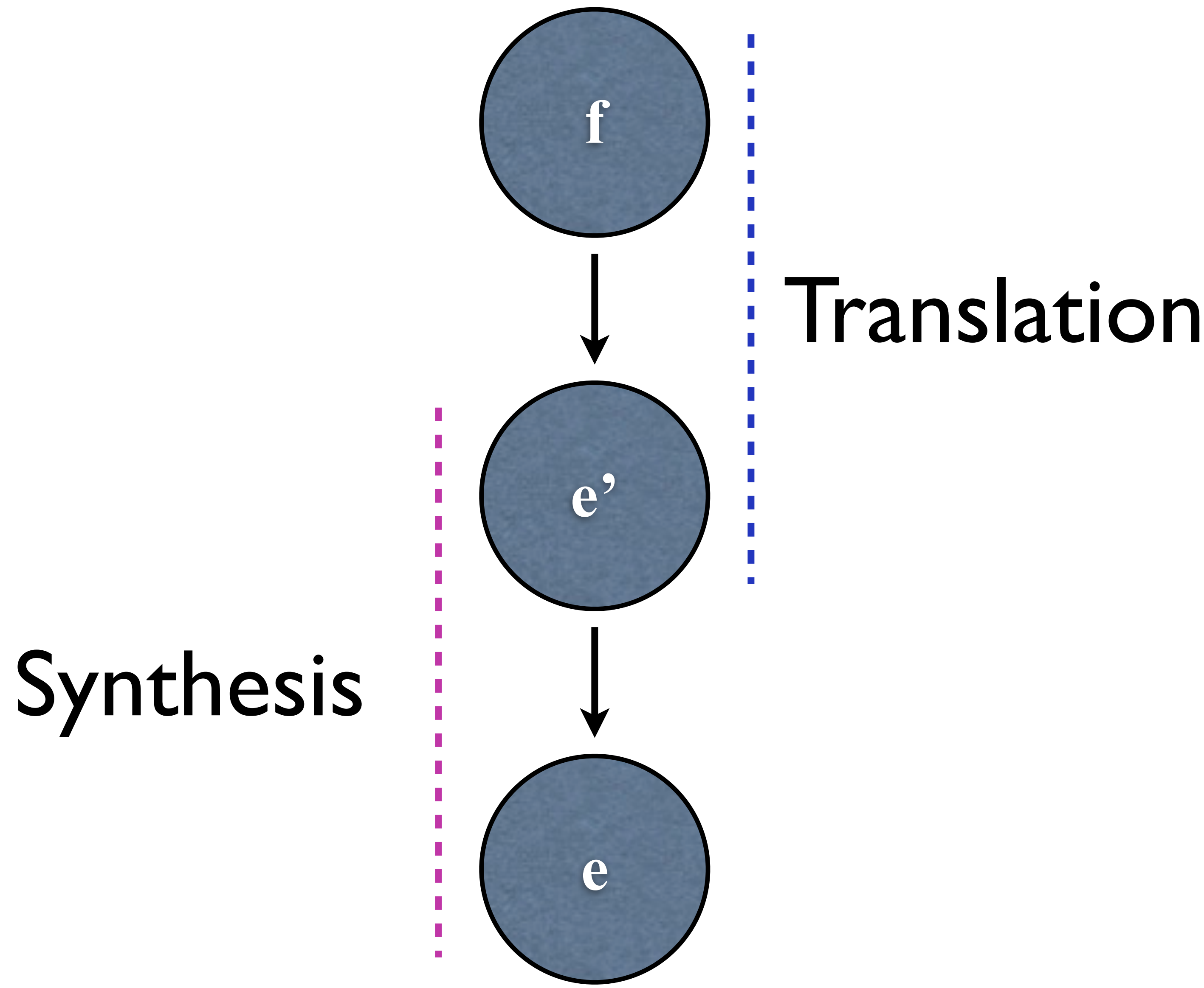


old



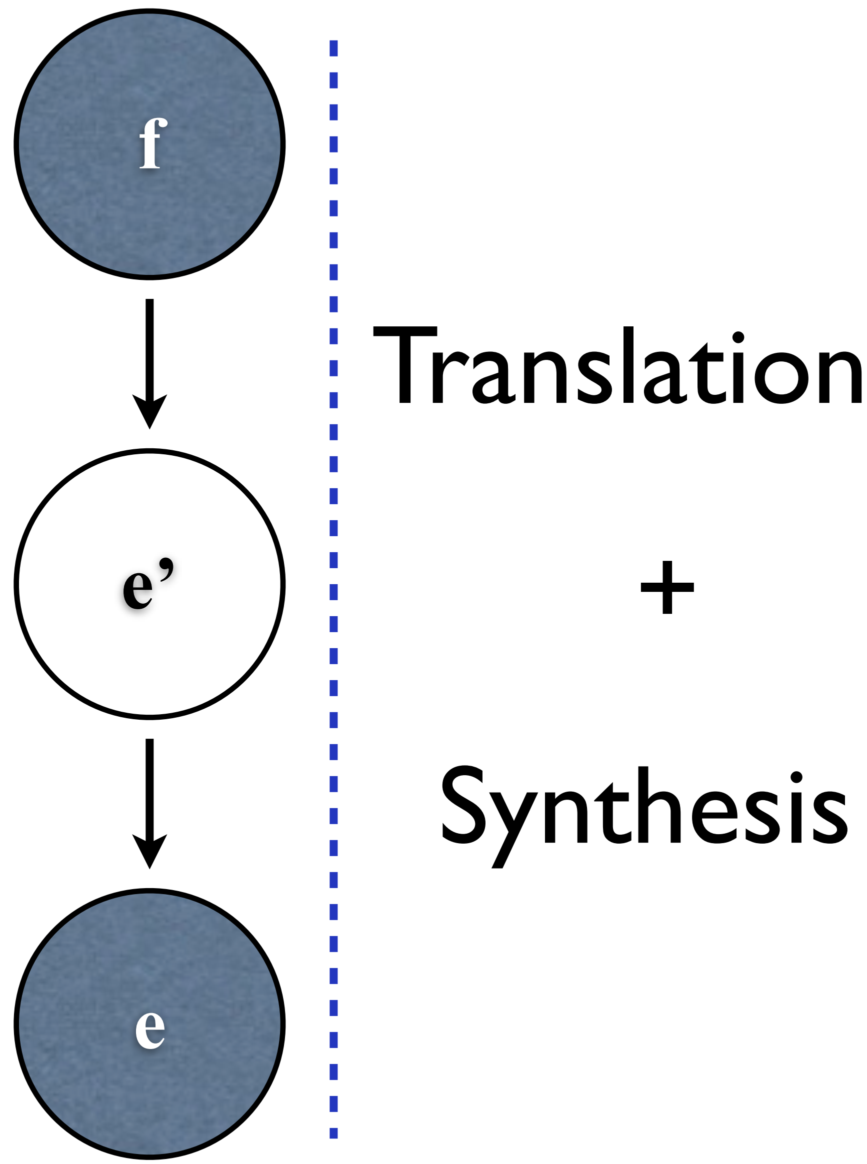
Modeling assumptions

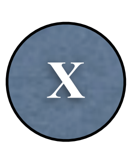
- Observed words are formed by an *unobserved* process that concatenates a stem **α** and a suffix **β**, yielding **αβ**
- A source word should have only a few translations **αβ**
- translate into only a few stems **α**
- The suffix **β** occurs many times, with many different stems
- **β** may be null
- **β** will have a maximum length of r
- Once a word has been translated into some inflected form, that *inflected form*, its *stem*, and its *suffix* should be more likely ("rich get richer")



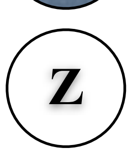
X Observed during training

Z Latent variable





Observed during training



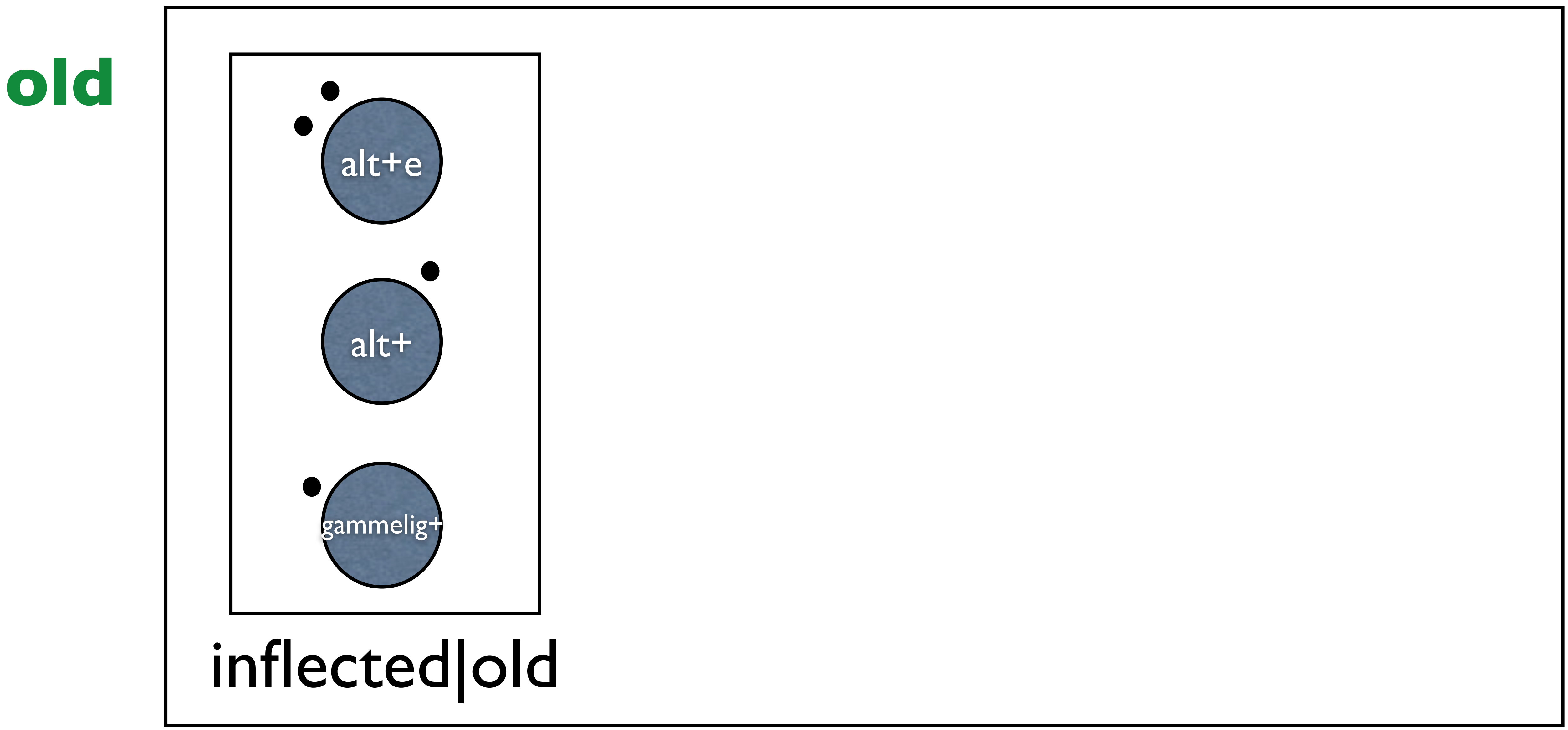
Latent variable

Task:

Translate the word **old**

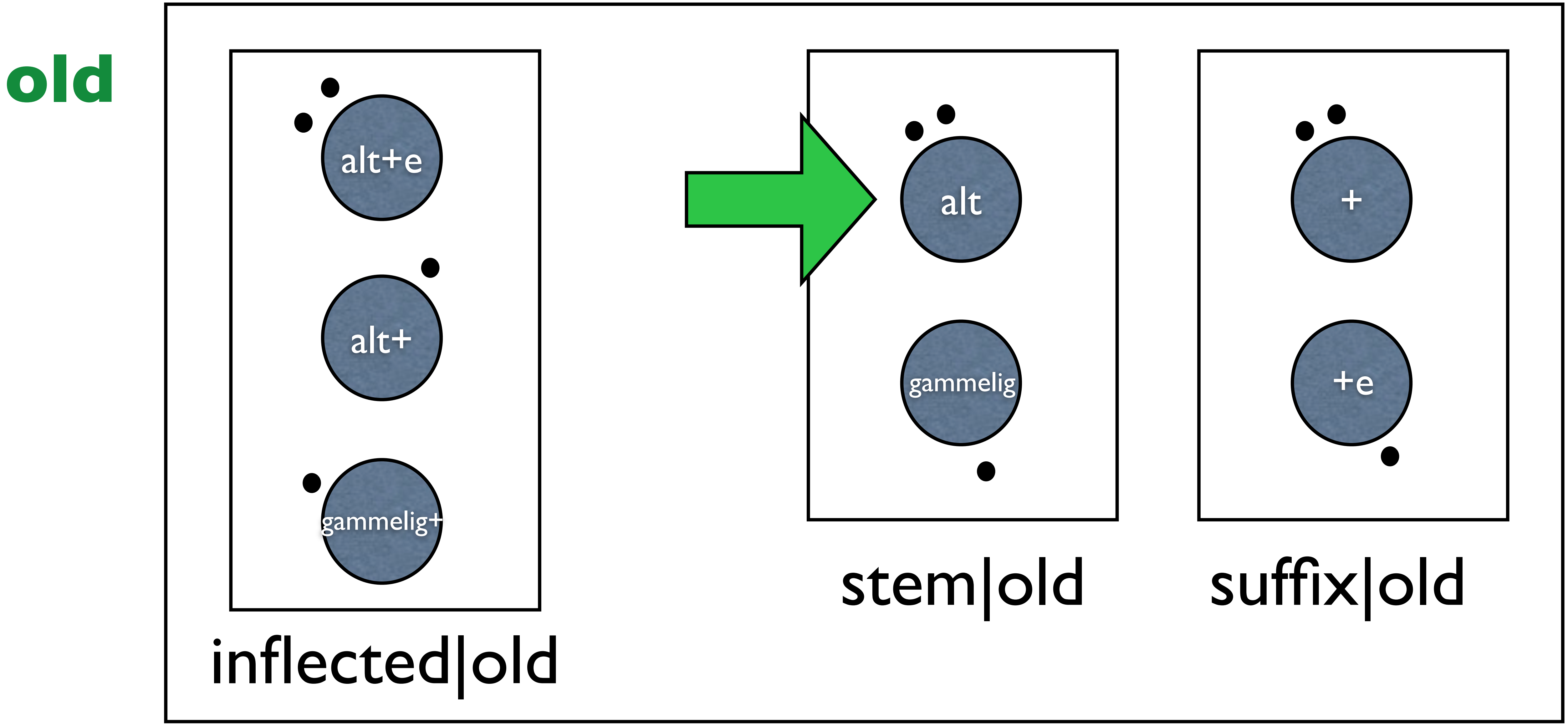
Task:

Translate the word **old**



Task:

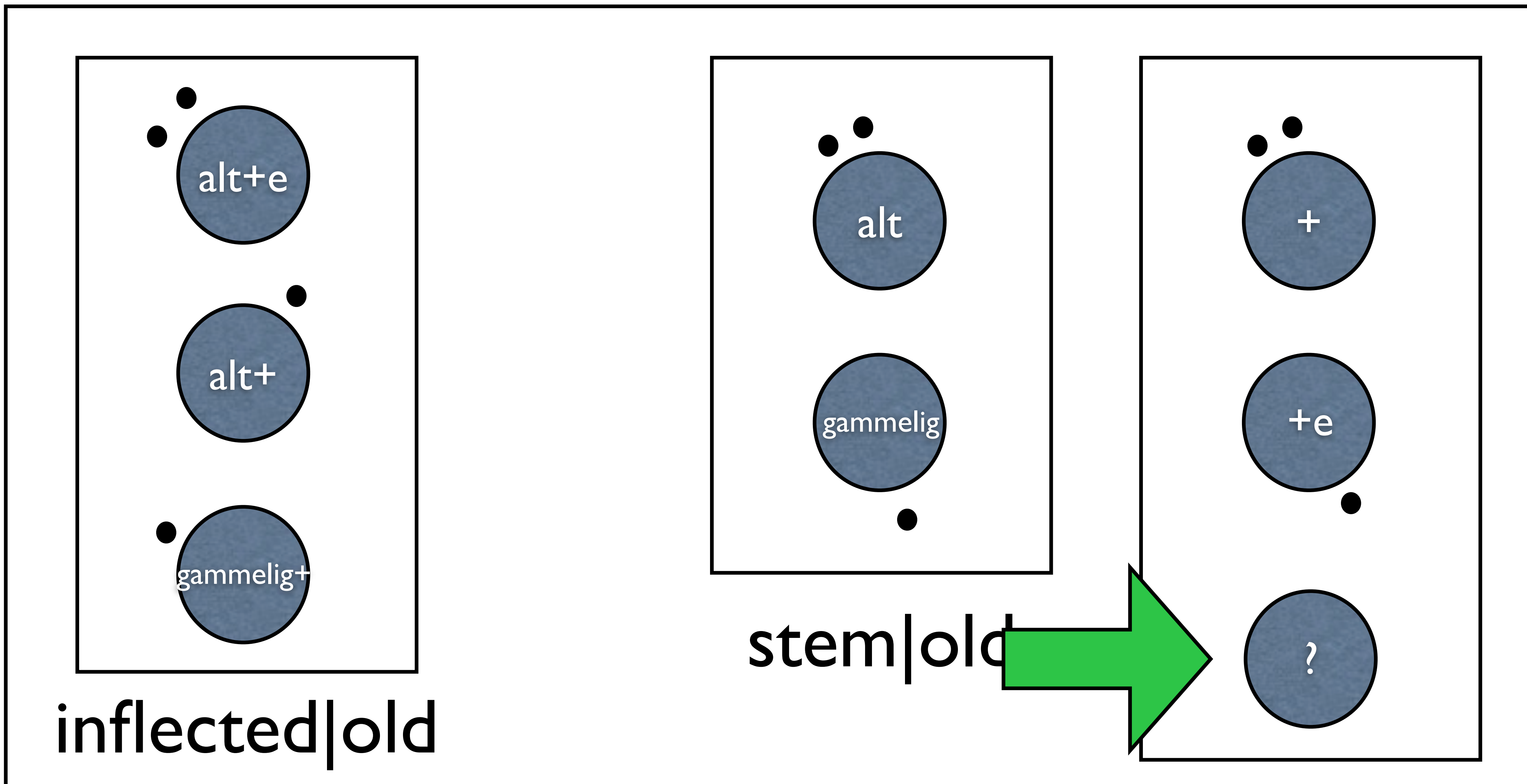
Translate the word **old** **alt**

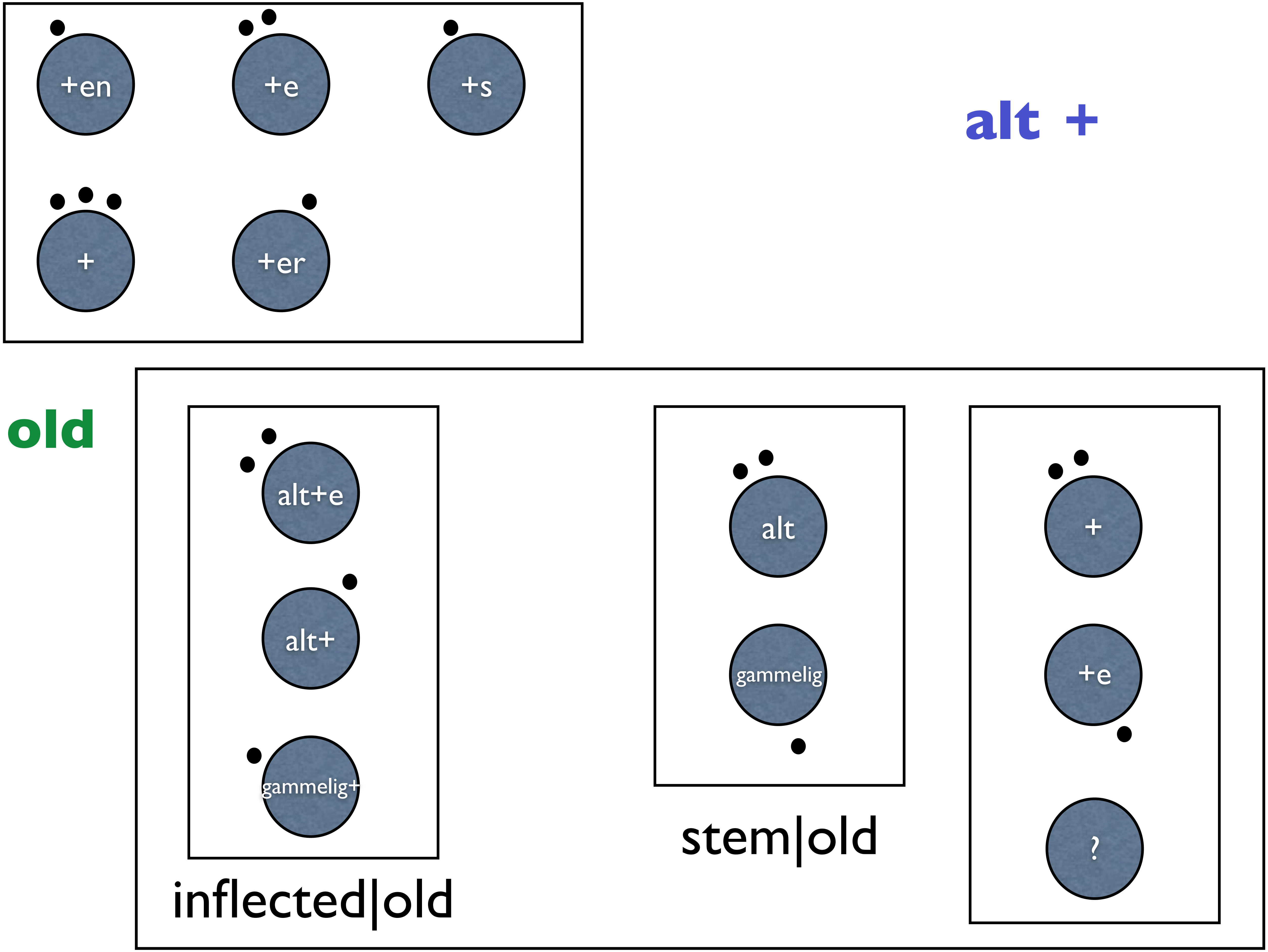


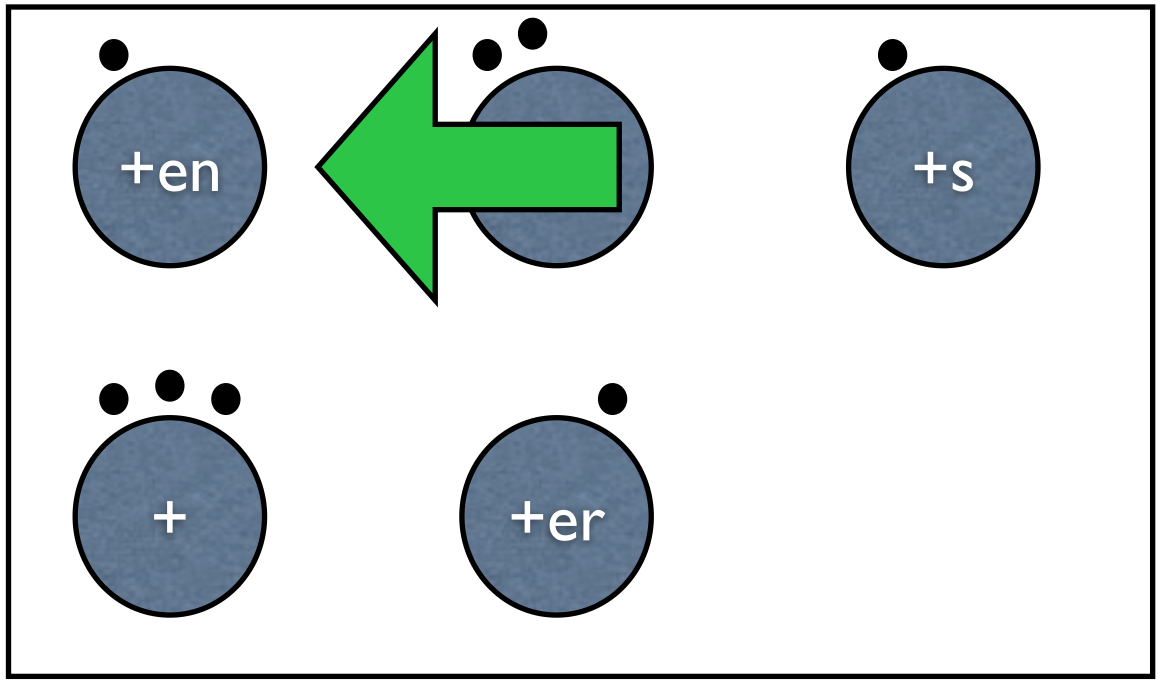
Task:

Translate the word **old** **alt +**

old

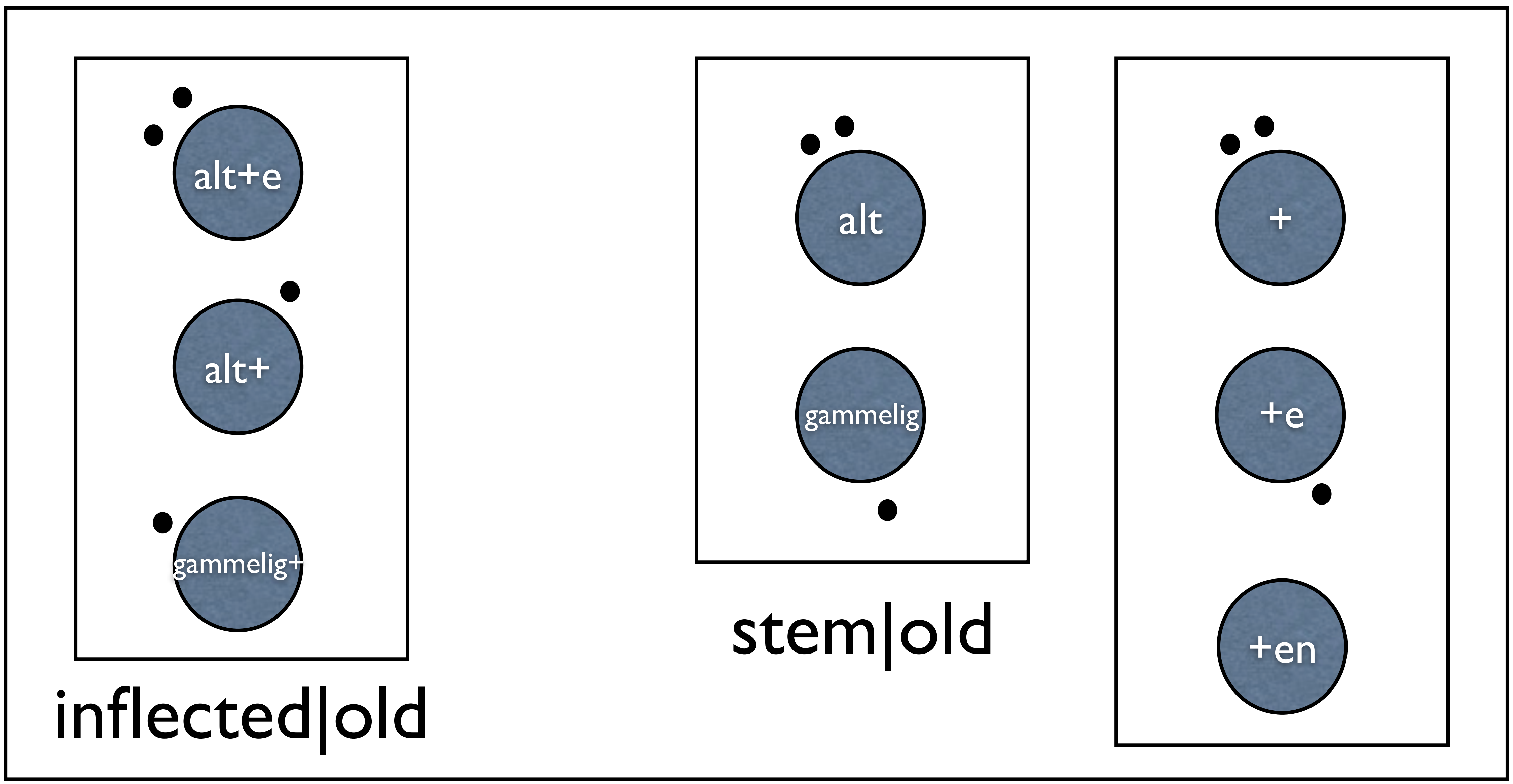



alt +
+en
+e
+s
+
+er
old
alt+e
alt+
gammelig+
inflected|old
alt
gammelig
stem|old
+
+e
?


+en
+s
+
+er

alt + en

old


alt+e
alt+
gammelig+
inflected|old
alt
gammelig
stem|old
+
+e
+en

Evaluation

- Given a parallel corpus, we can infer
  - The MAP alignment
  - The MAP **segmentation** of each target word into <stem+suffix>

Alignment Evaluation

		AER
Model 1 - EM	f|e	43.3
Model 1 - HPYP	f|e	**37.5**
Model 1 - EM	e|f	38.4
Model 1 - HPYP	e|f	**36.6**

English-French, 115k words, 447 sentences gold alignments.

Frequent suffixes

Suffix	Count
+∅	**20,837**
+s	**334**
+d	217
+e	156
+n	156
+y	130
+ed	**121**
+ing	**119**

Assessment

- Breaking the "lexical independence assumption" is computationally costly
  - The search space is much, much larger!
  - Dealing only with **inflectional morphology** simplifies the problems
- Sparse priors are crucial for avoiding degenerate solutions

In conclusion ...

Why don't we have integrated morphology?

Why don't we have integrated morphology?

Because we spend all our time working on English, which doesn't have much morphology!

Why don't we have integrated morphology?

- Translation with words is already hard: an n-word sentence has $n!$ permutations
- But, if you're looking at a sentence with m ***letters*** there are $m!$ permutations
  - Search is ... considerably harder
    - $m > n \quad \rightarrow \quad m! \ggg\gg n!$
  - Modeling is harder too
    - must also support all these permutations!

Take away messages

- Morphology matters for MT
- Probabilistic models are a great fit for the uncertainty involved
- Breaking the lexical independence assumption is hard

Thank you!

Toda!

$krAF!

https://github.com/redpony/cdec/

$$
\begin{aligned}
n &\sim \mathrm{Poisson}(\lambda)\\
a_i &\sim \mathrm{Uniform}(1/|\mathbf{f}|)\\
e_i \mid f_{a_i} &\sim T_{f_{a_i}}\\
T_{f_{a_i}} \mid a, b, \mathrm{M} &\sim \mathrm{PYP}(a, b, \mathrm{M}(\cdot \mid f_{a_i}))\\
\mathrm{M}(e = \alpha + \beta \mid f) &= G_f(\alpha) \times H_f(\beta)\\
G_f \mid a, b, f, \mathrm{P}_0 &\sim \mathrm{PYP}(a, b, \mathrm{P}_0(\cdot))\\
H_f \mid a, b, f, \mathrm{H}_0 &\sim \mathrm{PYP}(a, b, \mathrm{H}_0(\cdot))\\
H_0 \mid a, b, \mathrm{Q}_0 &\sim \mathrm{PYP}(a, b, \mathrm{Q}_0(\cdot))\\
\mathrm{P}_0(\alpha; p) &= \frac{p^{|\beta|}}{|V|^{|\beta|}} \times (1 - p)\\
\mathrm{Q}_0(\beta; r) &= \frac{1}{(|V| \times r)^{|\beta|}}
\end{aligned}
$$