Translating into Morphologically Rich Languages

Chris Dyer, CMU

November 22, 2013

The Problem

- A Swahili verb can have **hundreds** of different inflectional forms
- A Malagasy noun will have **dozens** of different forms

Challenge: How can we hope to learn all these forms, *particularly in a low resource scenario*?

nilimuona	paka
1SG+5CL_OBJ+***see***+PAST	**cat**

I saw ***the*** cat???

I saw a	
I saw	a cat
I see	cat

nilimuona	paka
1SG+5CL_OBJ+***see***+PAST	**cat**

kusoma (V)

“to read”

	+1p +sg	+3p +sg	+1p +pl
+pres	ninasoma	anasoma	tunasoma
	“I am reading”	“he/she is reading”	“we are reading”
+past	nilisoma	alisoma	tulisoma
	“I read”	“he/she read”	“he/she read”

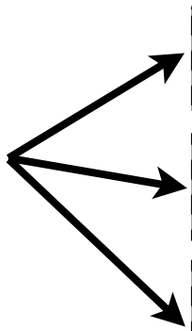
		+object:cl7 e.g., kitabu="the book"	+object:cl8 e.g., vitabu="the books"
+pres	ninasoma	ninakisoma	ninavisoma
	"I am reading"	"I am reading it"	"I am reading them"
+past	nilisoma	nilikisoma	nilivisoma
	"I read"	"I read it"	"I read them"

Observation: inflectional labels have *structure*.

I saw a	
I saw	a cat
I see	cat

nilimuona	paka
1SG+5CL_OBJ+***see***+PAST	**cat**

Synthetic phrases



I saw the cat	
I see the	
I saw the	
I saw a	the cat
I saw	a cat
I see	cat

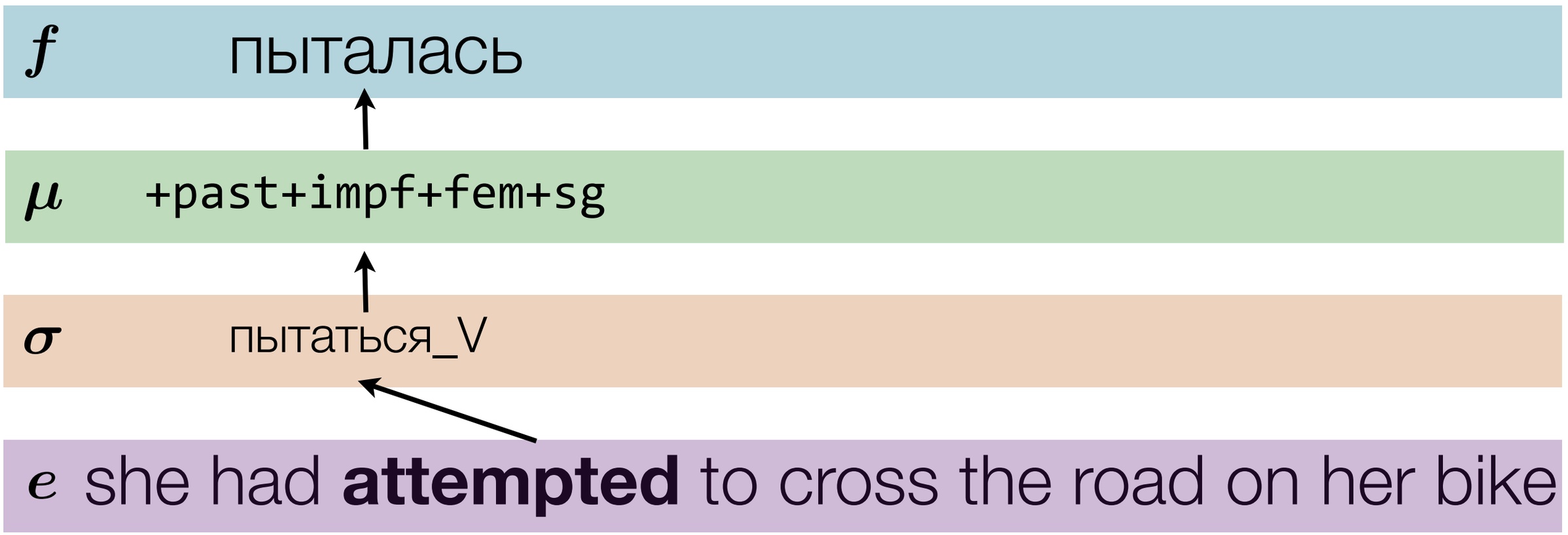
nilimuona	paka
1SG+5CL_OBJ+***see***+PAST	**cat**

The Morphological Process

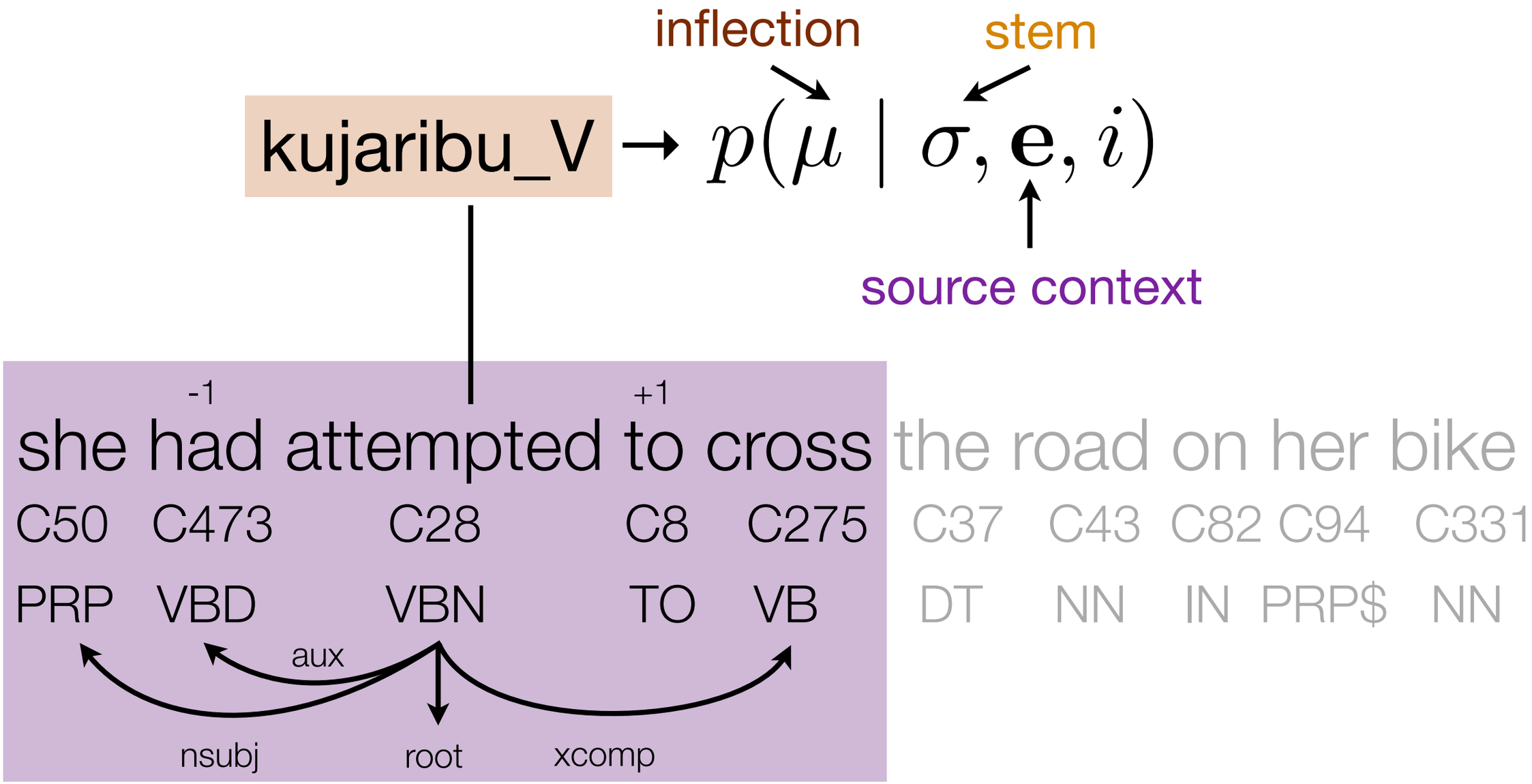
stem inflection target word

$$(\sigma : \text{пытаться\_V}) \star (\mu : \texttt{+past+impf+fem+sg}) = (f : \text{пытатьсь})$$

$$\sum_{\sigma \star \mu = f} \Big[\underbrace{p(\sigma \mid e_i)}_{\text{gen. stem}} \times \underbrace{p(\mu \mid \sigma, \boldsymbol{e}, i)}_{\text{gen. inflection}} \Big] = p(f \mid \boldsymbol{e}, i)$$

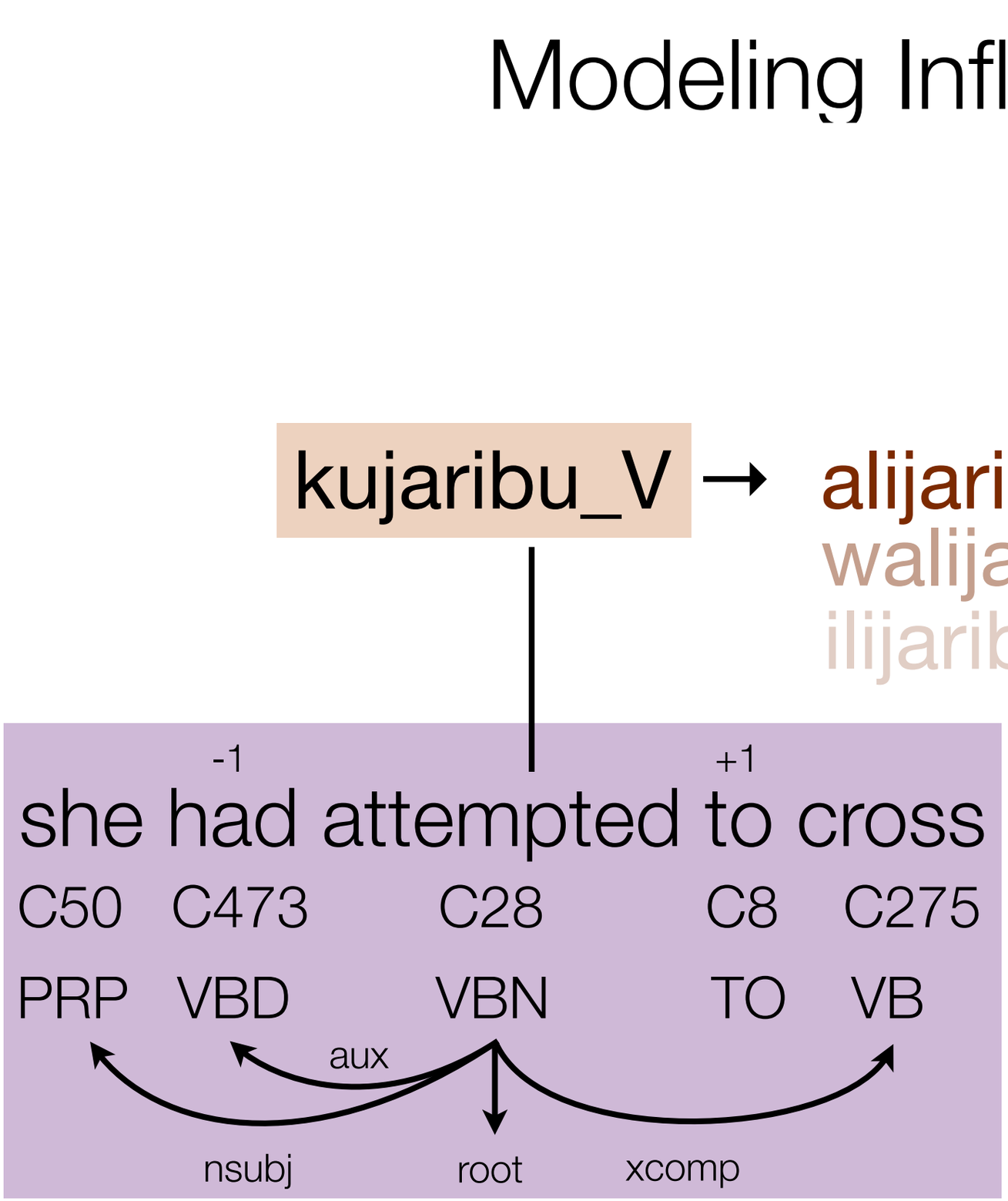


Modeling Inflection



Modeling Inflection

kujaribu_V → alijaribu
walijaribu
ilijaribu
} ranked list



Modeling Inflection

inflection stem

$$p(\mu \mid \sigma, \mathbf{e}, i) = \frac{u(\mu, \mathbf{e}, i)}{\sum_{\mu' \in \Omega_\sigma} u(\mu', \mathbf{e}, i)}$$

source context

$$u(\mu, \mathbf{e}, i) = \exp\left[\boldsymbol{\varphi}(\mathbf{e}, i)^\top \mathbf{W} \boldsymbol{\psi}(\mu)\right]$$

$\boldsymbol{\varphi}(\mathbf{e}, i)$: (features) source context

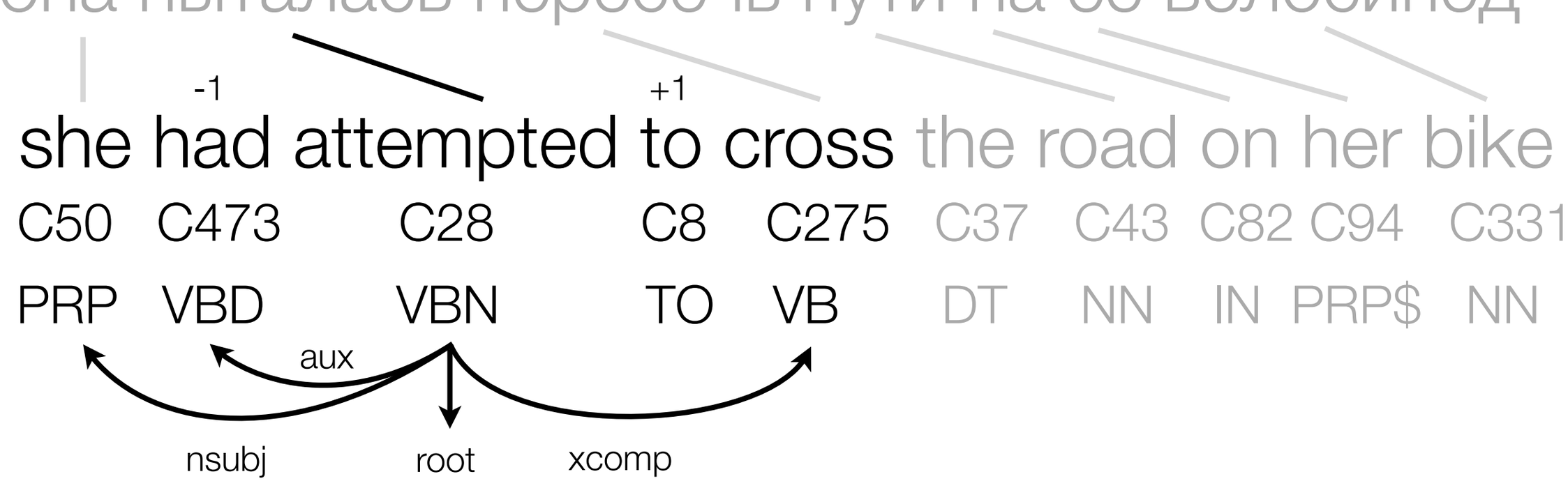
$\boldsymbol{\psi}(\mu)$: (features) target morphology

$\mathbf{W}$: (parameters) map input to output

Source context features

$$\boldsymbol{\varphi}(\mathbf{e}, i)$$

она пыталась пересечь пути на ее велосипед



Output Morphology Features

$$\boldsymbol{\psi}(\mu)$$

ni+li+STEM →

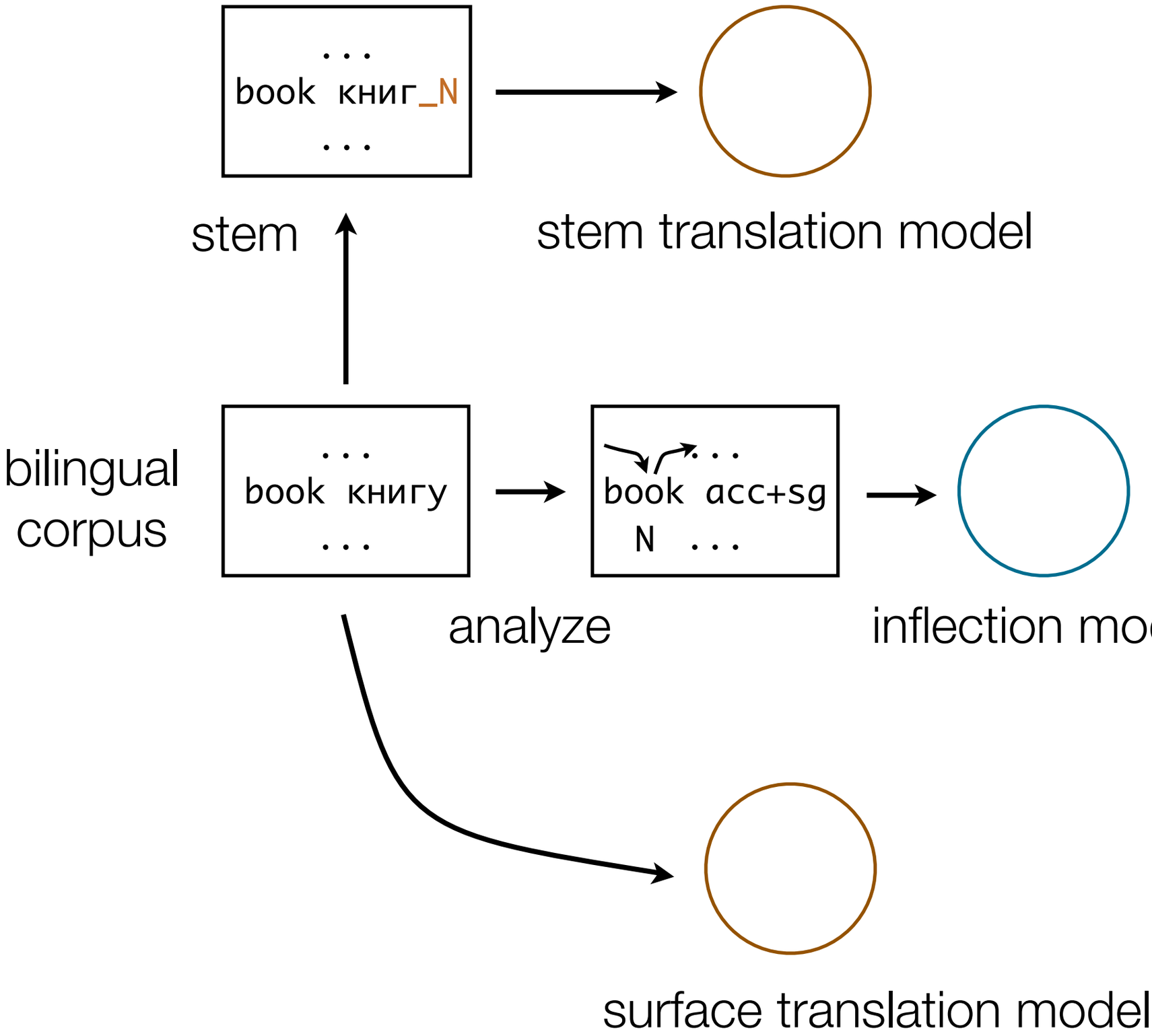
Prefix[-1][li]:1
Prefix[-2][ni]:1

ni+me+STEM →

Prefix[-1][me]:1
Prefix[-2][ni]:1

Learning the Model Parameters

- Obtain the set of all inflections for each lemma from *monolingual* data
- Extract training instances from word aligned *bilingual* data
- Train model parameters $\mathbf{w}$


...
book книг_N
...
stem
stem translation model
bilingual corpus
...
book книгу
...
...
book acc+sg
N ...
analyze
inflection model
surface translation model

inflection model

stem translation model

surface translation model

inflection model

“With his book”

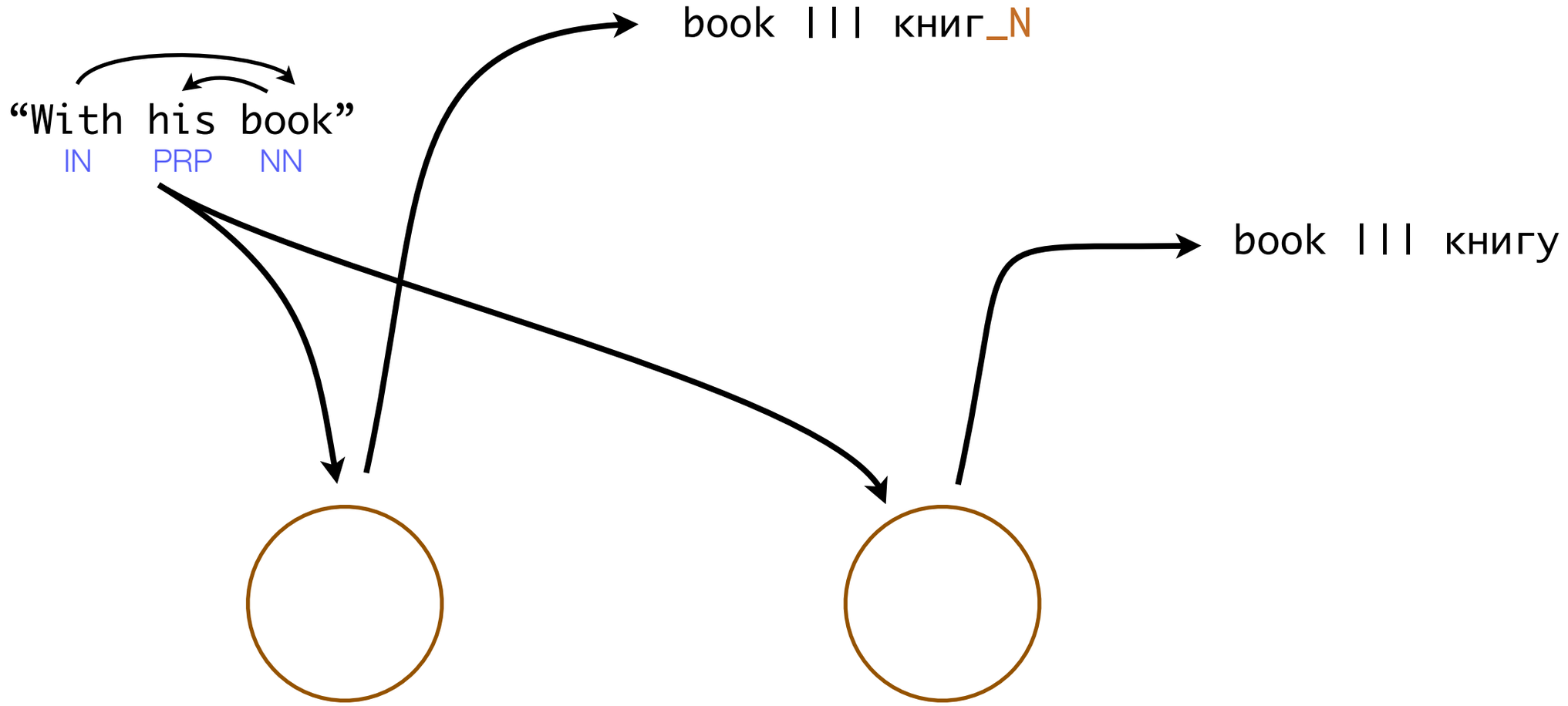
IN PRP NN

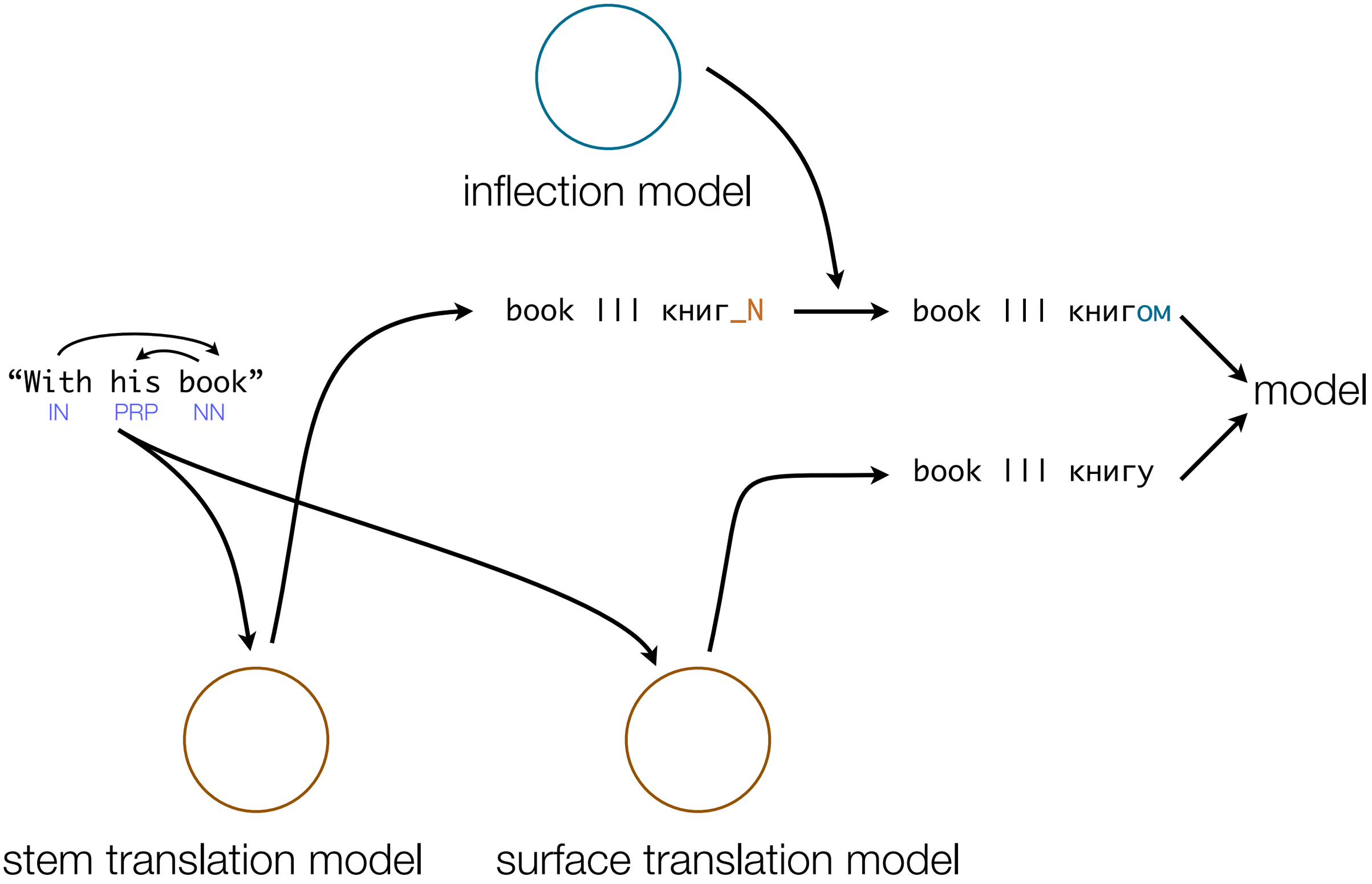
book III книгу

stem translation model

surface translation model

inflection model




inflection model
book ||| книг_N
book ||| книгом
model
book ||| книгу
“With his book”
IN PRP NN
stem translation model
surface translation model

Translation results

	en → Russian	en → Swahili	en → Hebrew
Baseline (Status: 2012)	14.7±0.1	18.3±0.1	15.8±0.3
+ morphology	**16.7±0.1**	**19.0±0.1**	**17.6±0.1**

$$(\sigma : \text{question}) \star (\mu : \textbf{+PL}) =$$

questions ?

What is learned?

Prefix *li* (gloss: PAST)

`source=VBD` `source=VBN`

Prefix *nita* (gloss: 1P-SING + FUTURE)

`child(aux)` `child(nsubj)=I`

Prefix *ana* (gloss: 3P-SING + PRESENT)

`source=VBZ`

Prefix *wa* (gloss: 3P-PLURAL)

`before=they` `child(nsubj)=NNS`

Suffix *tu* (gloss: 1P-PLURAL)

`child(nsubj)=she` `before=she`

Prefix *ha* (gloss: NEG)

`source=no` `after=not`

What is learned?

Prefix *li* (gloss: PAST)

`source=VBD source=VBN`

Prefix *nita* (gloss: 1P-SING + FUTURE)

`child(aux) child(nsubj)=I`

Prefix *ana* (gloss: 3P-SING + PRESENT)

`source=VBZ`

Prefix *wa* (gloss: 3P-PLURAL)

`before=they child(nsubj)=NNS`

Suffix *tu* (gloss: 1P-PLURAL)

`child(nsubj)=she before=she`

Prefix *ha* (gloss: NEG)

`source=no after=not`

What is learned?

Prefix *li* (gloss: PAST)

`source=VBD source=VBN`

Prefix *nita* (gloss: 1P-SING + FUTURE)

`child(aux) child(nsubj)=I`

Prefix *ana* (gloss: 3P-SING + PRESENT)

`source=VBZ`

Prefix *wa* (gloss: 3P-PLURAL)

`before=they child(nsubj)=NNS`

Suffix *tu* (gloss: 1P-PLURAL)

`child(nsubj)=she before=she`

Prefix *ha* (gloss: NEG)

`source=no after=not`