Cross-Lingual Learning for Natural Language Syntax

Dipanjan Das

July 26, 2014

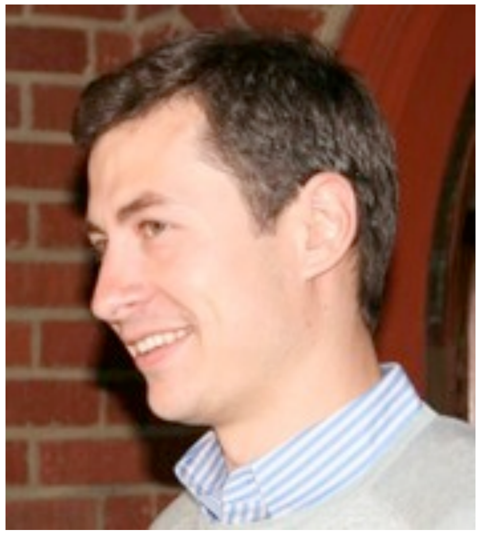
Lisbon Machine Learning School

Collaborators

Kuzman Ganchev
Google

Ryan McDonald
Google

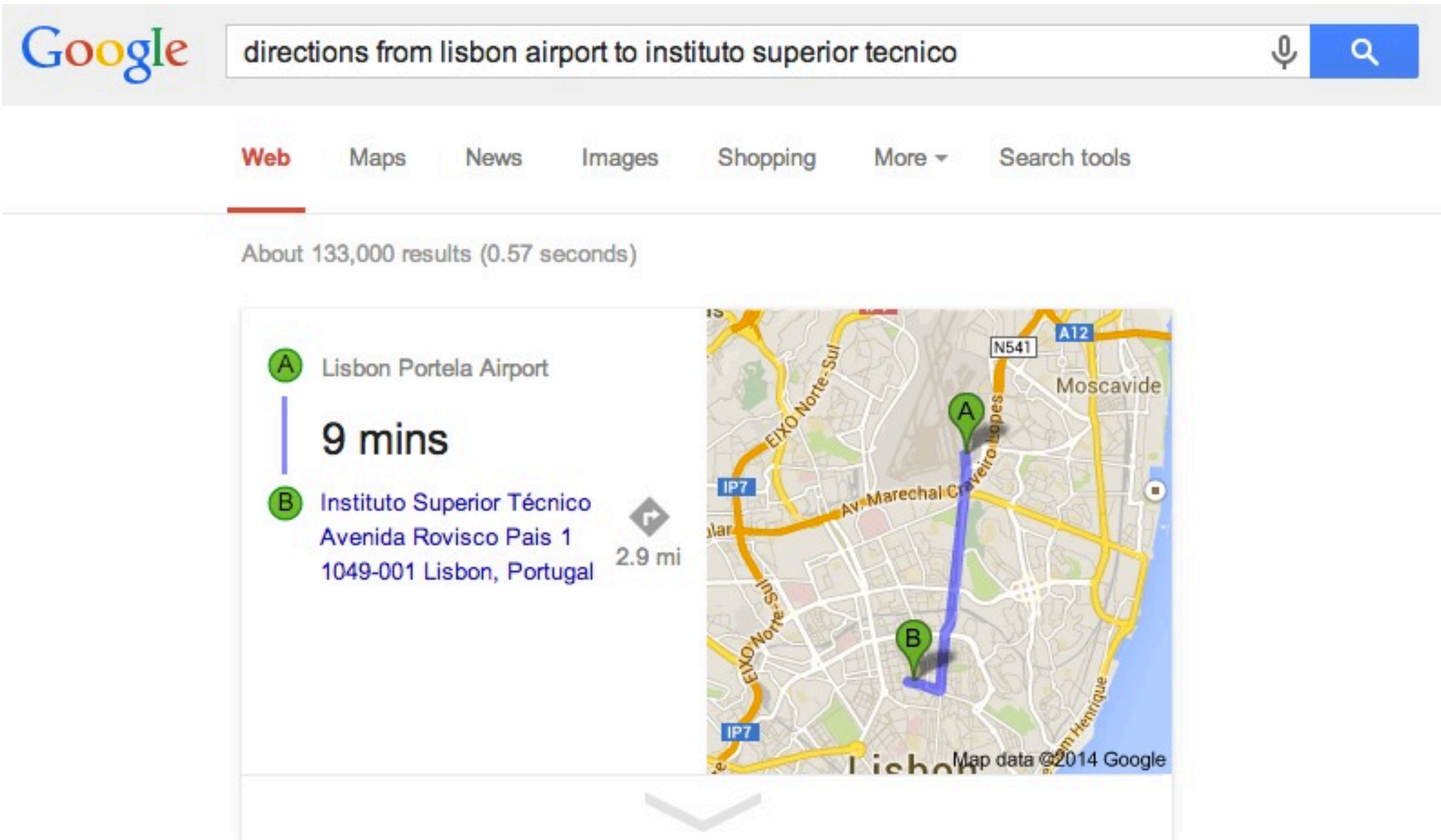
Joakim Nivre
Uppsala University



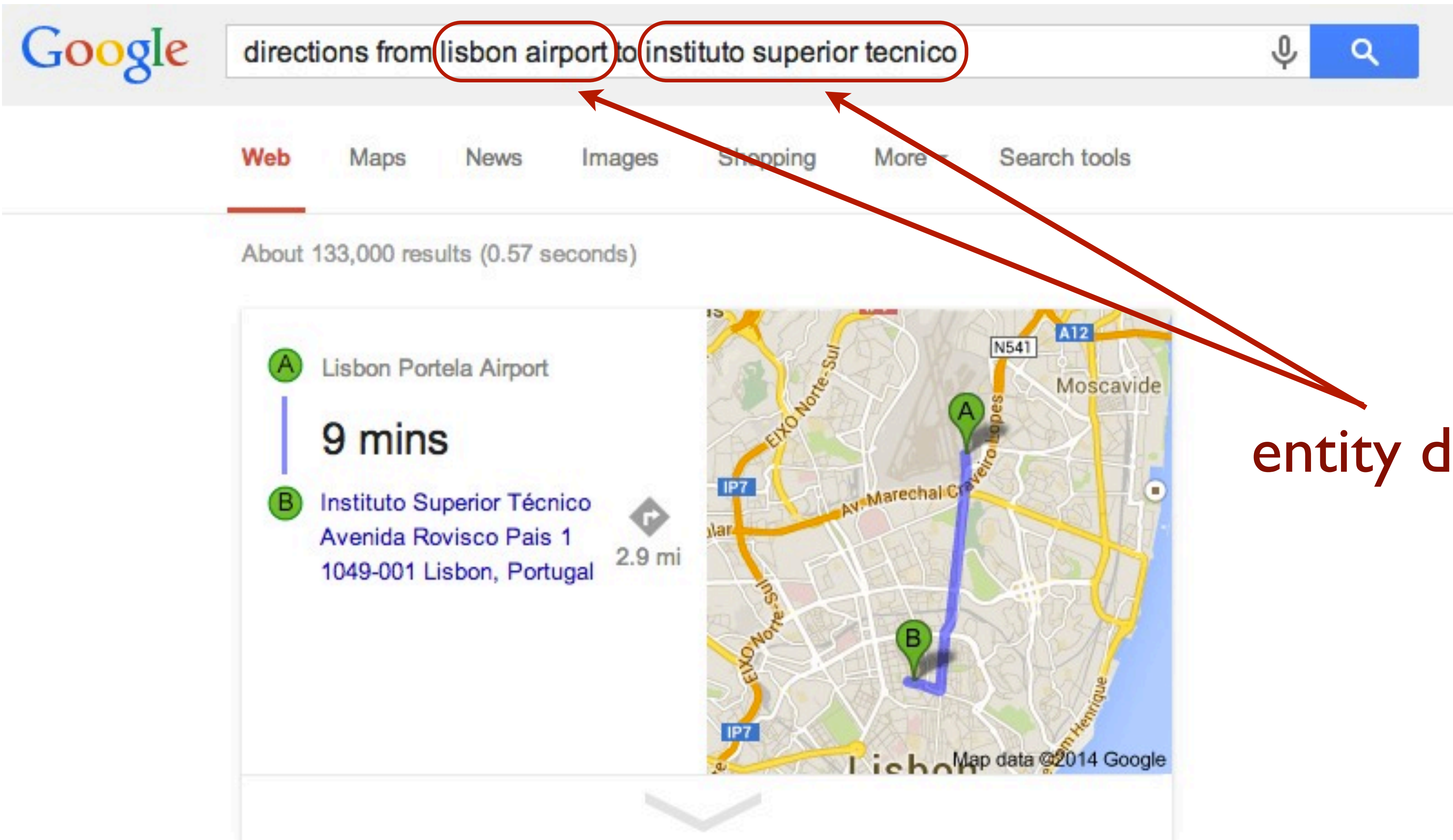
Slav Petrov
Google

Oscar Täckström
Google

Natural Language Processing in Practice

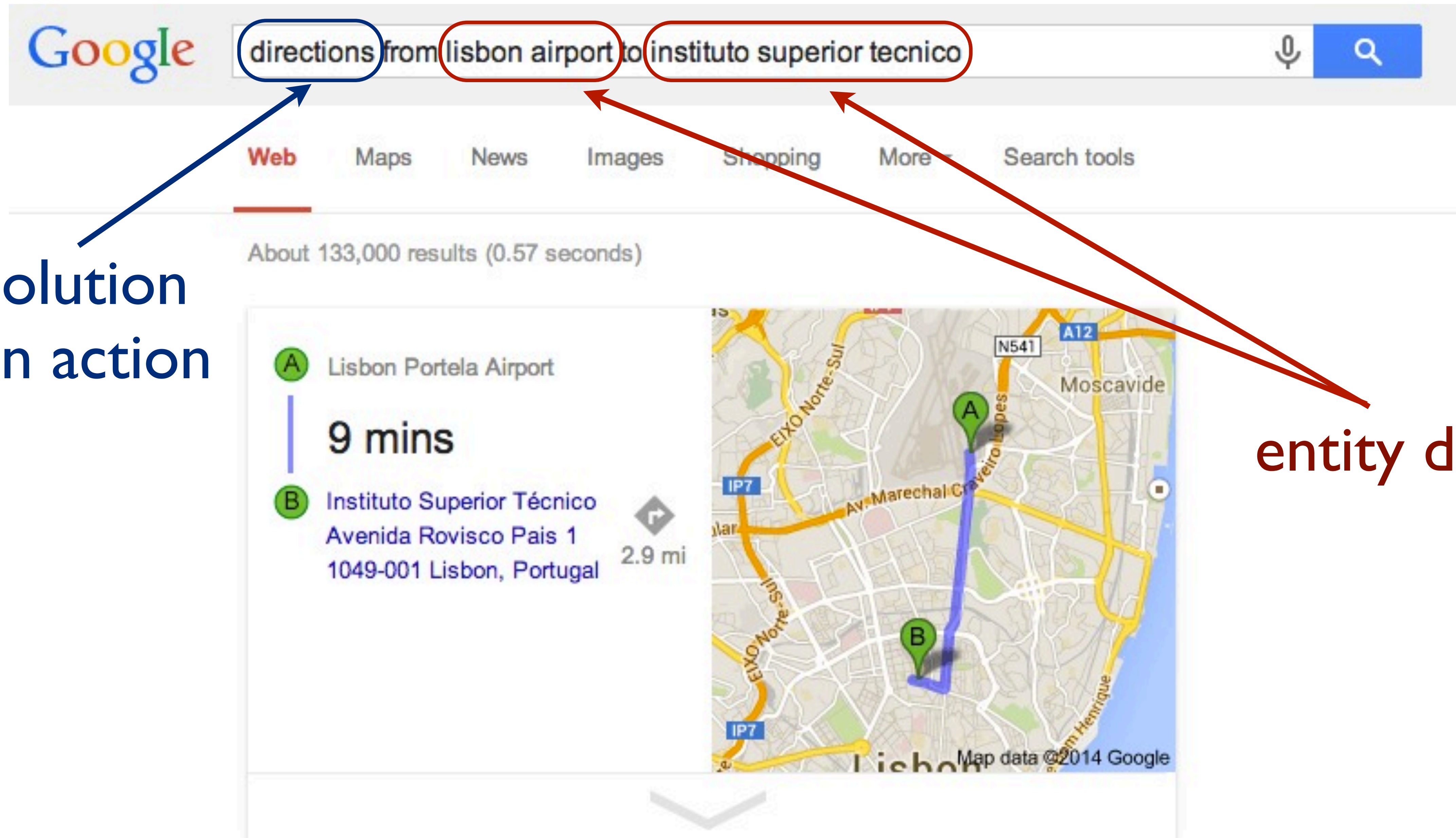


Natural Language Processing in Practice Google



entity detection

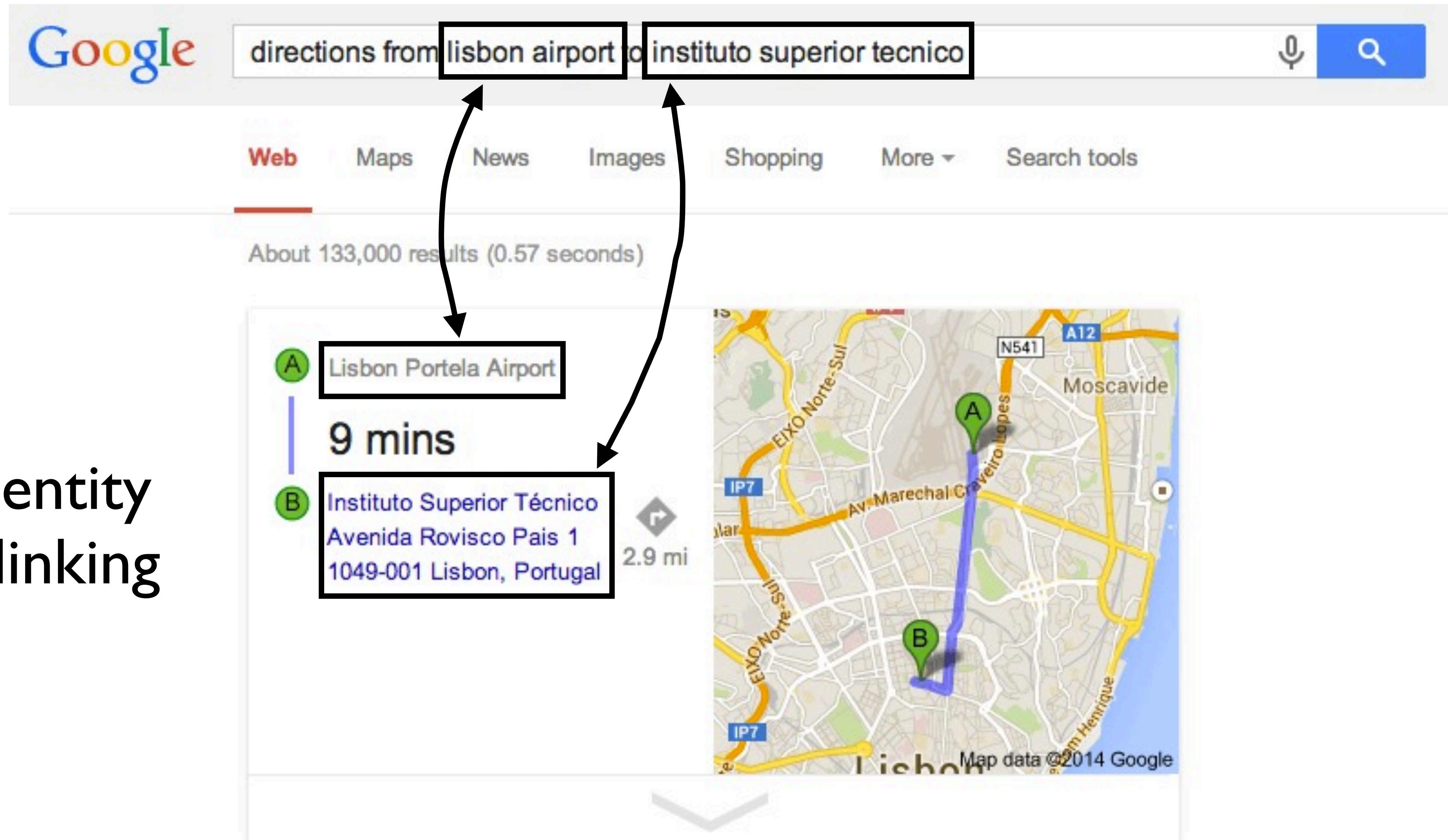
Natural Language Processing in Practice Google



resolution to an action

entity detection

Natural Language Processing in Practice Google



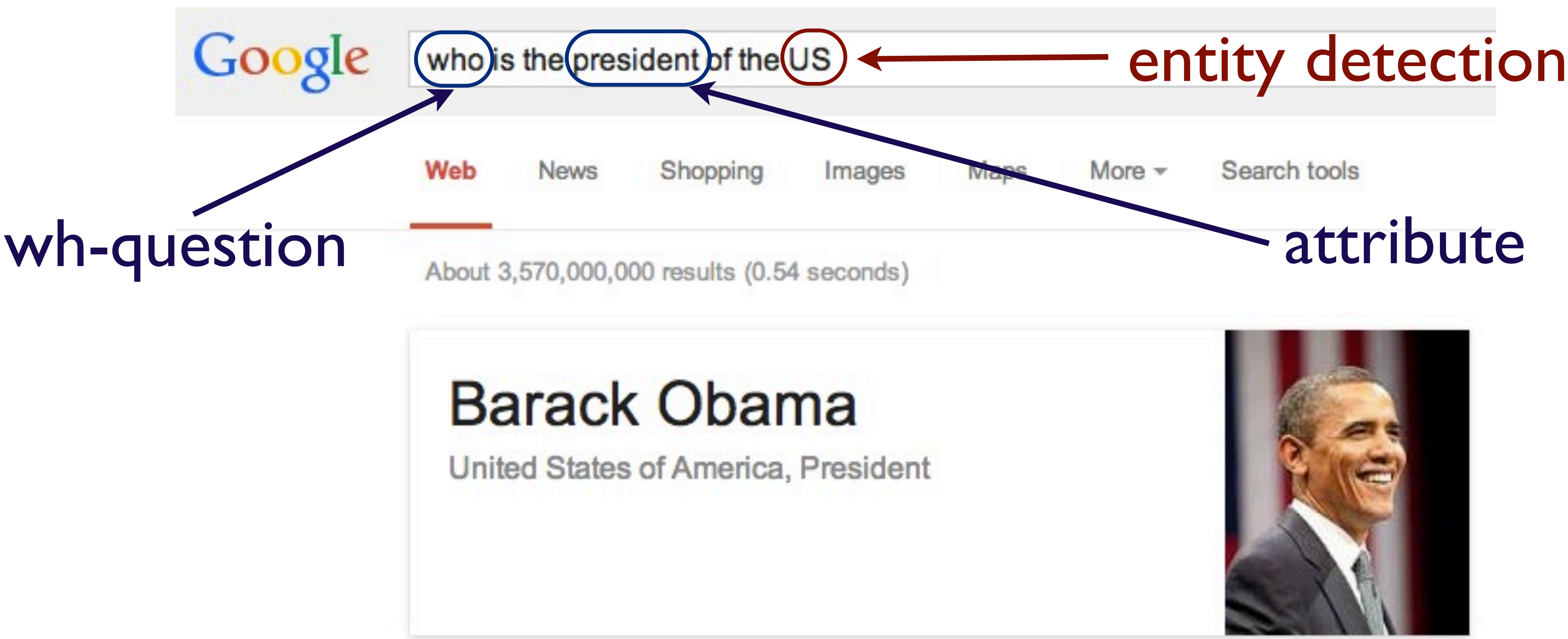
entity linking

Natural Language Processing in Practice Google

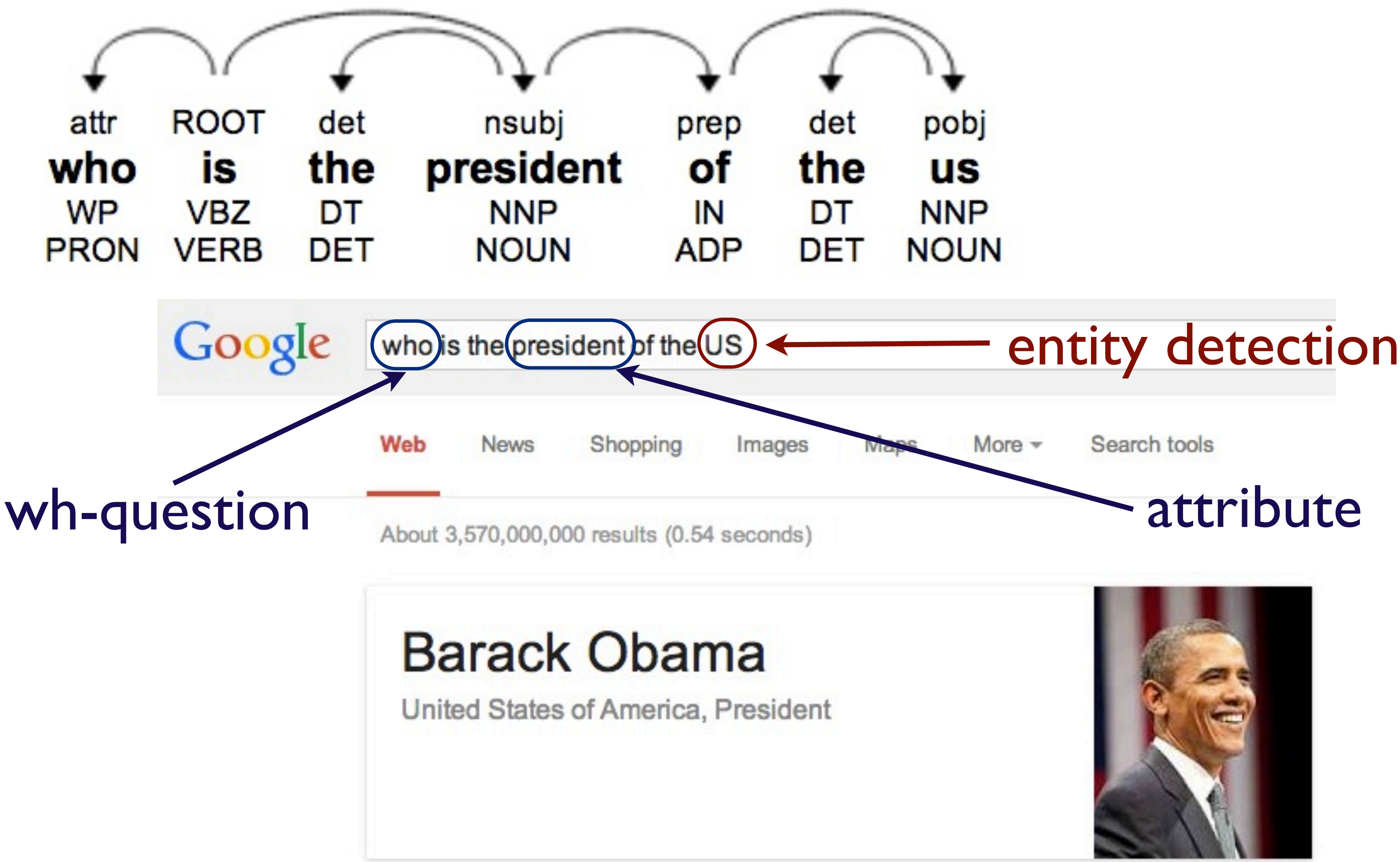
Natural Language Processing in Practice Google

Natural Language Processing in Practice

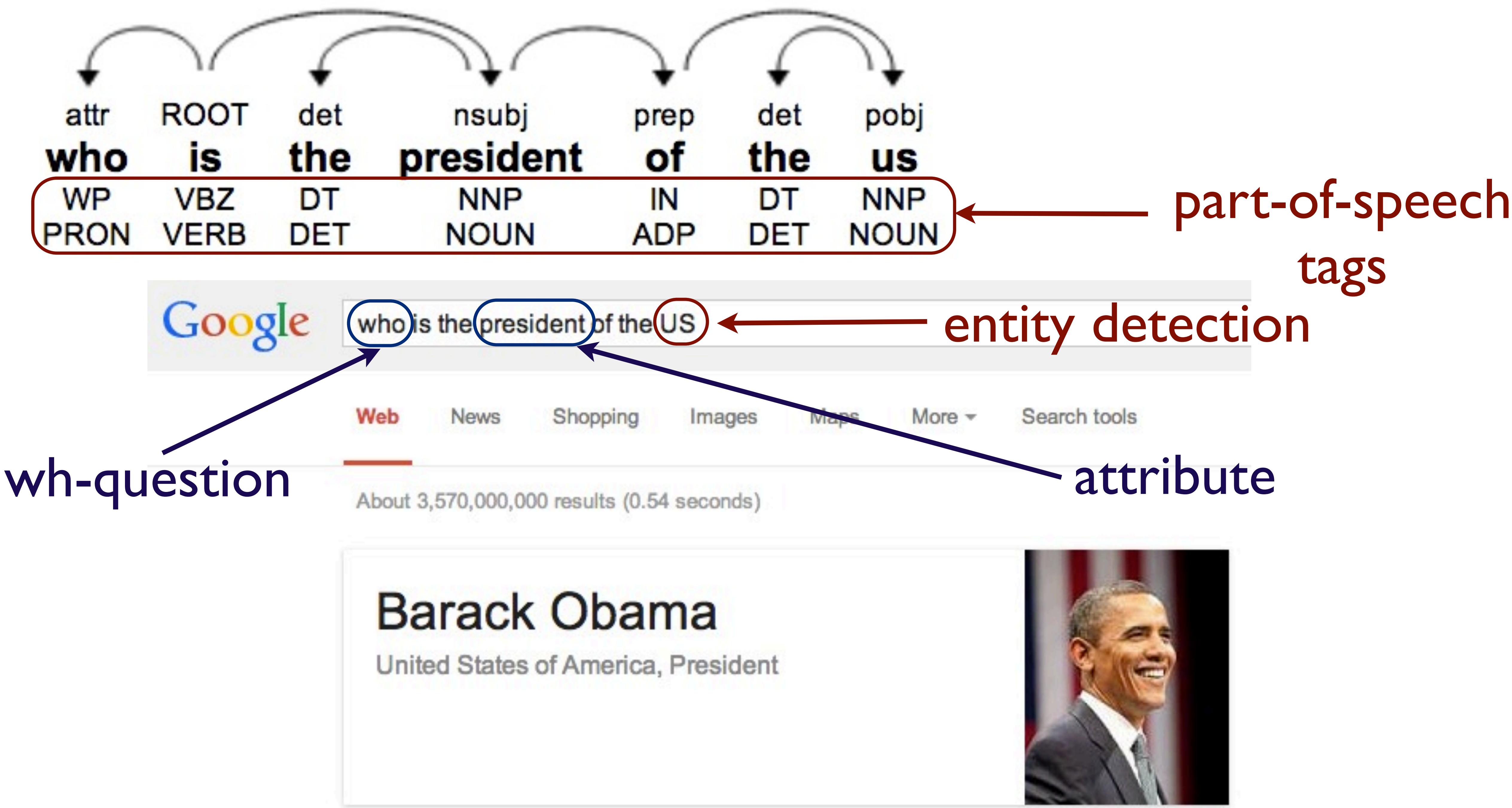
Natural Language Processing in Practice



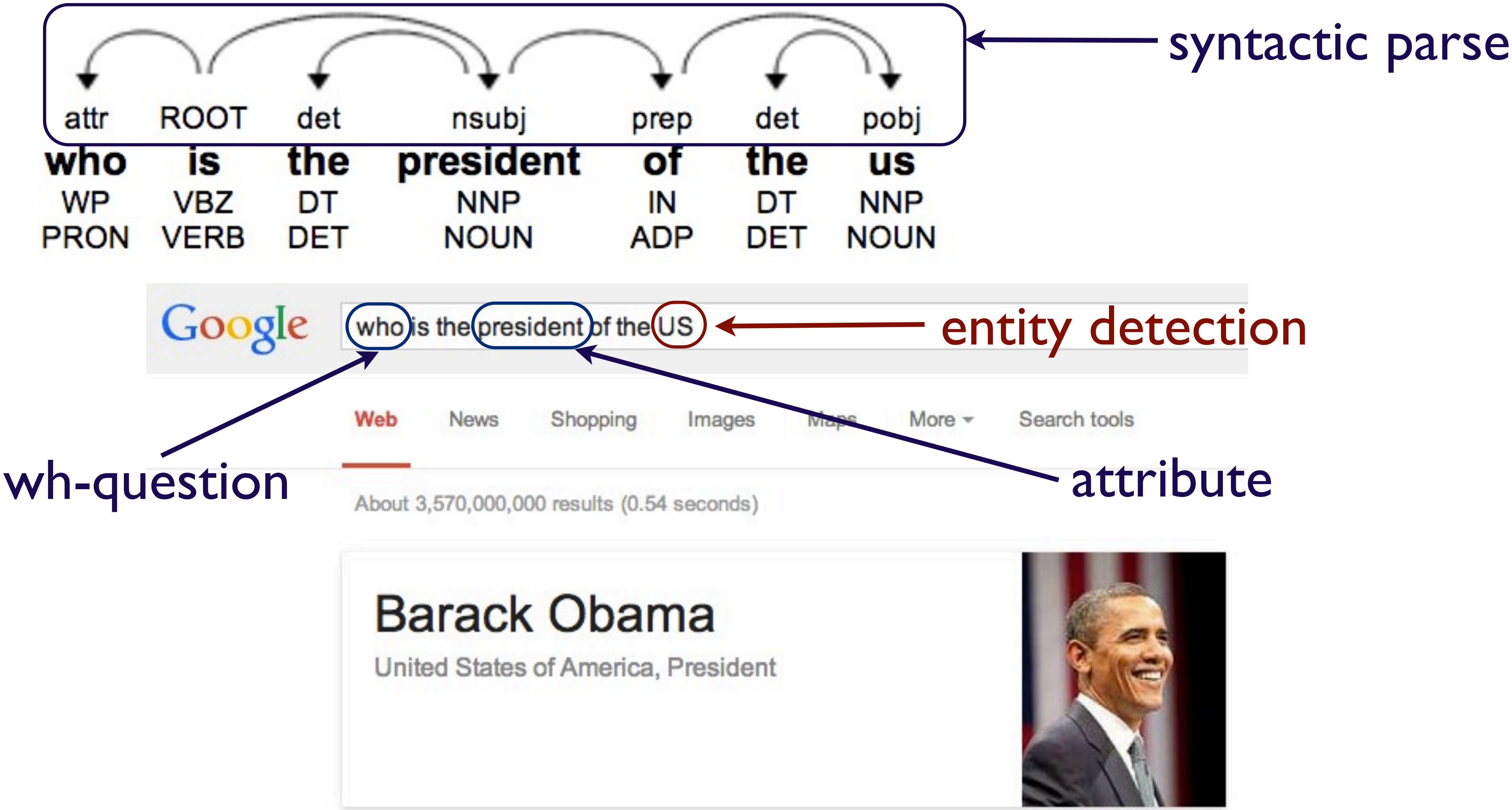
Natural Language Processing in Practice Google



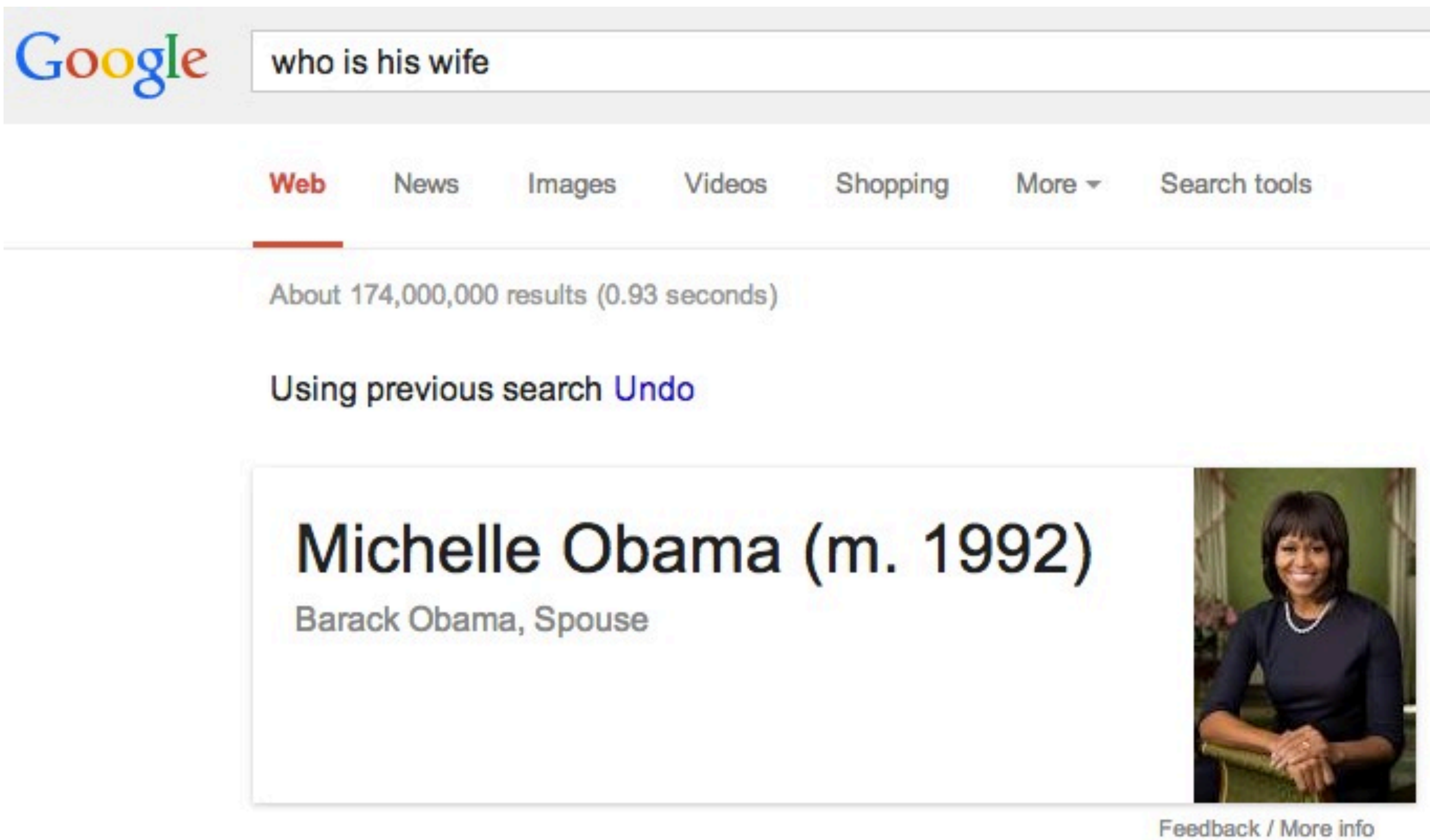
Natural Language Processing in Practice Google



Natural Language Processing in Practice Google

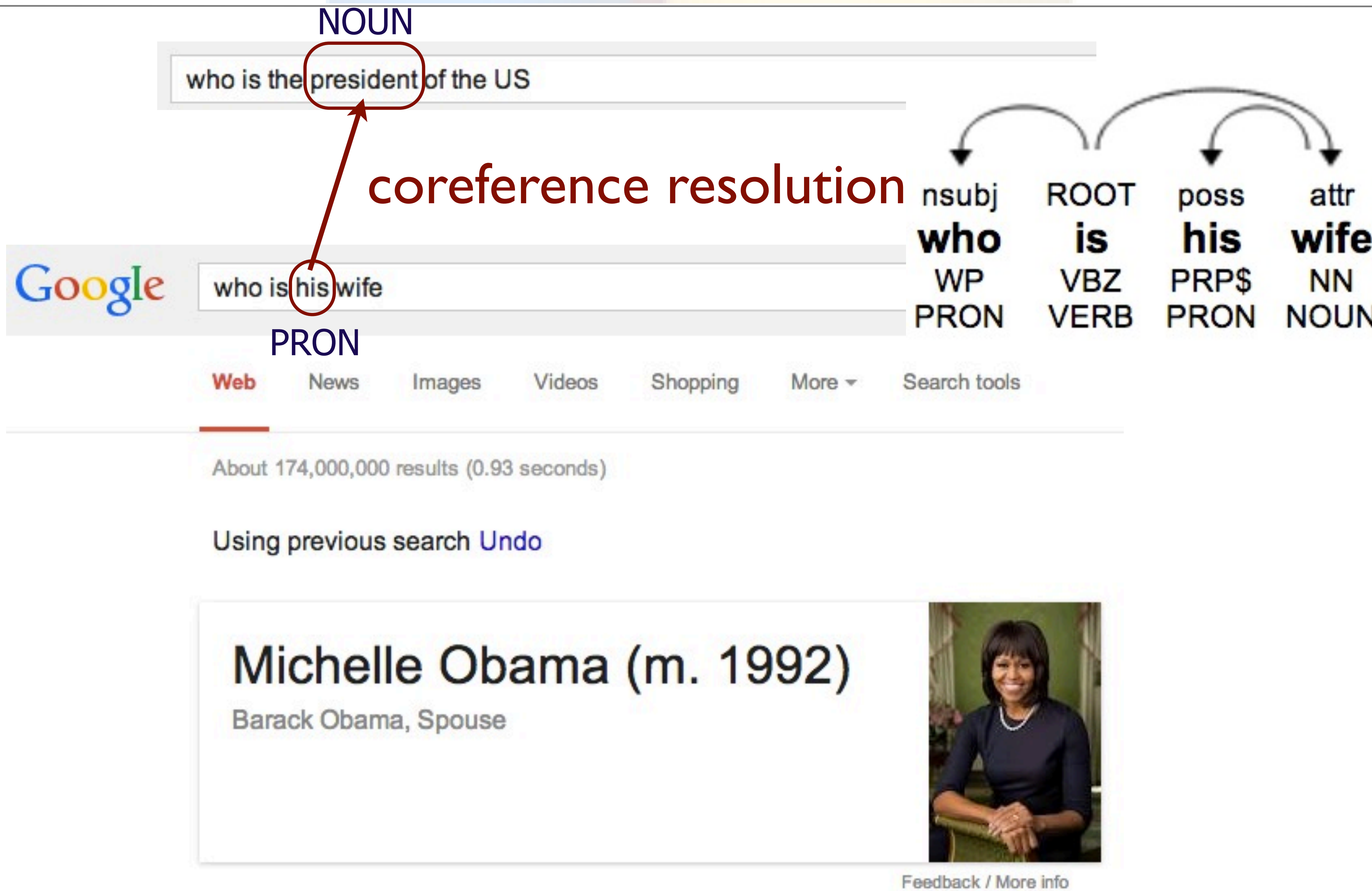


Natural Language Processing in Practice



Natural Language Processing in Practice Google

Natural Language Processing in Practice Google



Natural Language Processing in Practice Google

Natural Language Processing in Practice Google

who is his wife

Google when was she born

Web Images News Shopping Videos More Search tools

About 87,400,000 results (0.84 seconds)

Using previous search Undo

January 17, 1964 (age 50 years)

Michelle Obama, Date of birth

Natural Language Processing in Practice Google

who is his wife

coreference resolution

Google when was she born

Web Images News Shopping Videos More Search tools

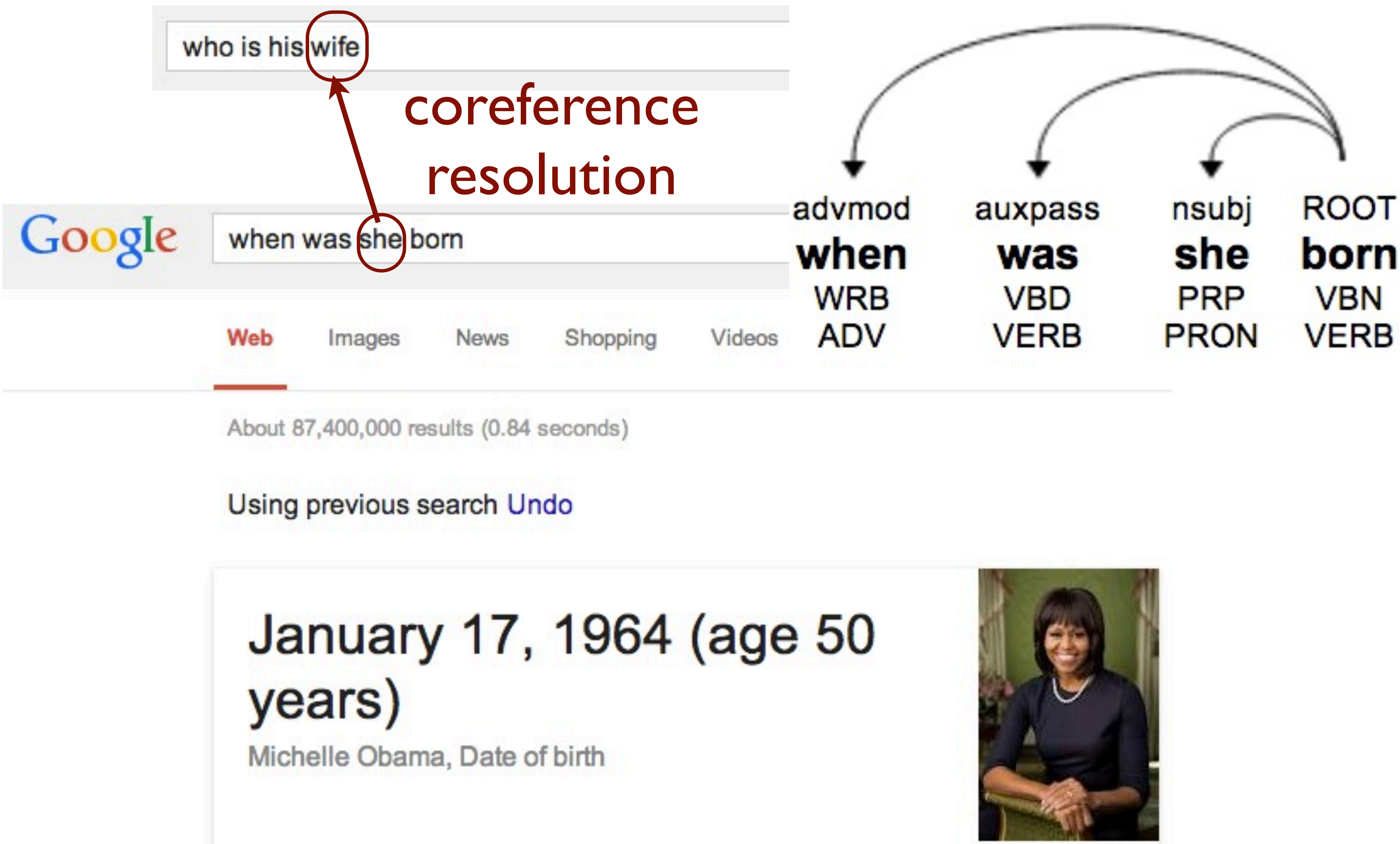
About 87,400,000 results (0.84 seconds)

Using previous search Undo

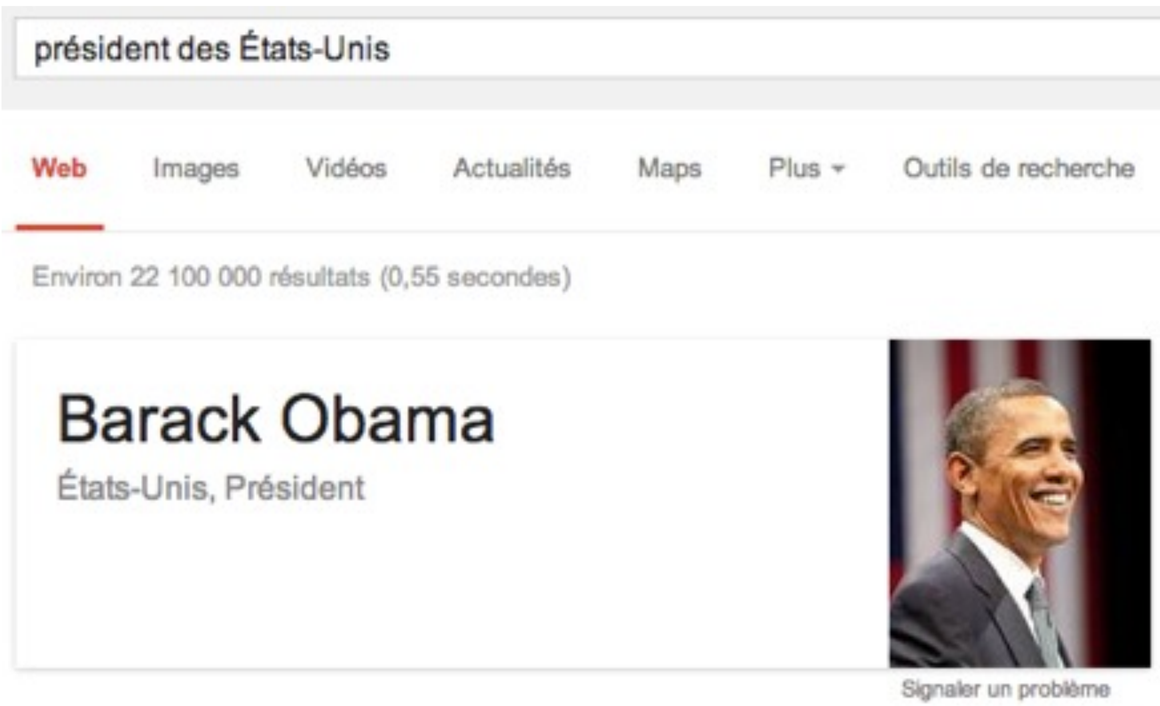
January 17, 1964 (age 50 years)

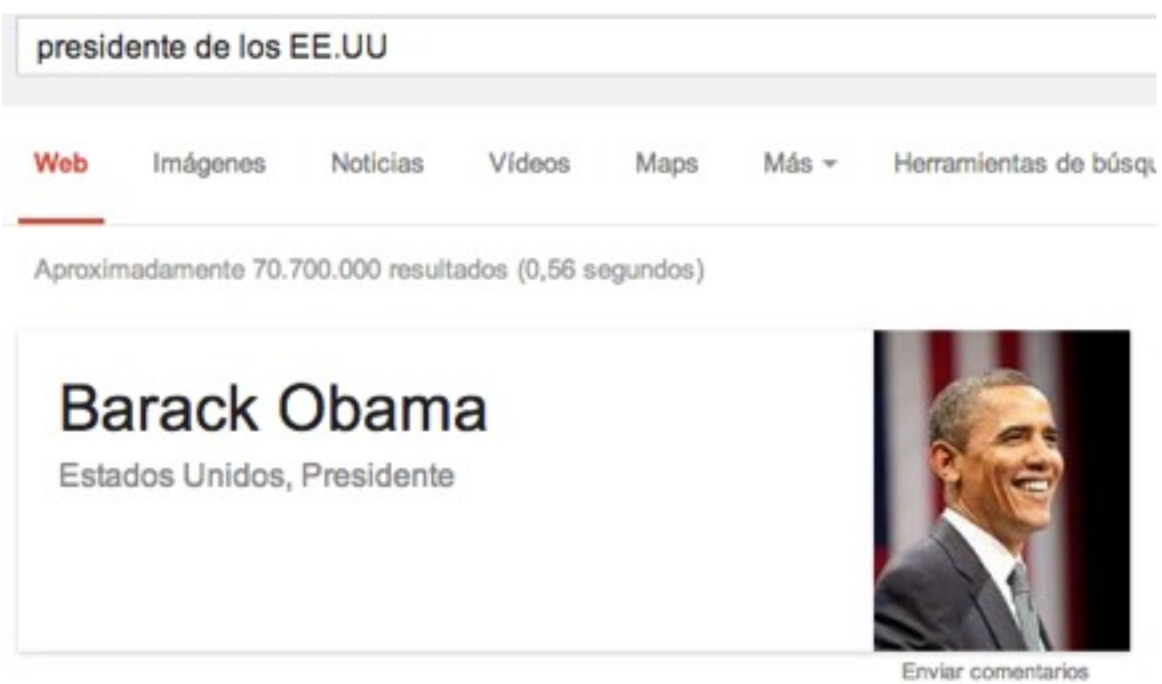
Michelle Obama, Date of birth

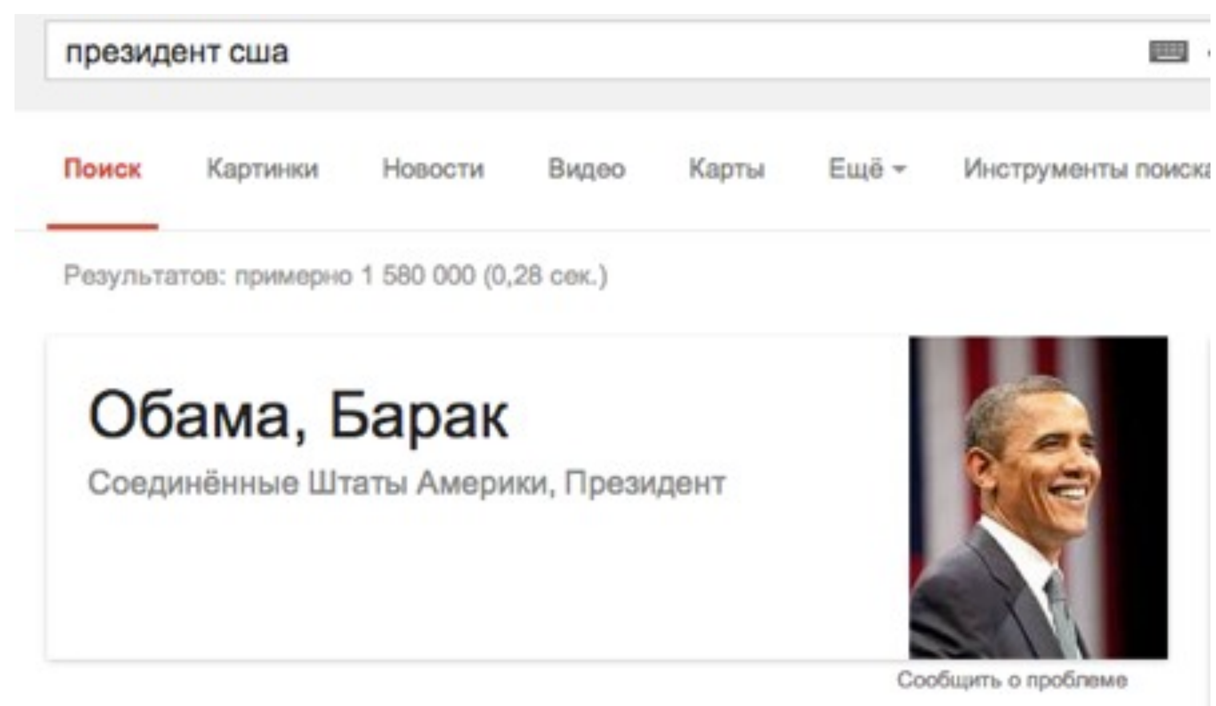
Natural Language Processing in Practice



Natural Language Processing in Practice Google

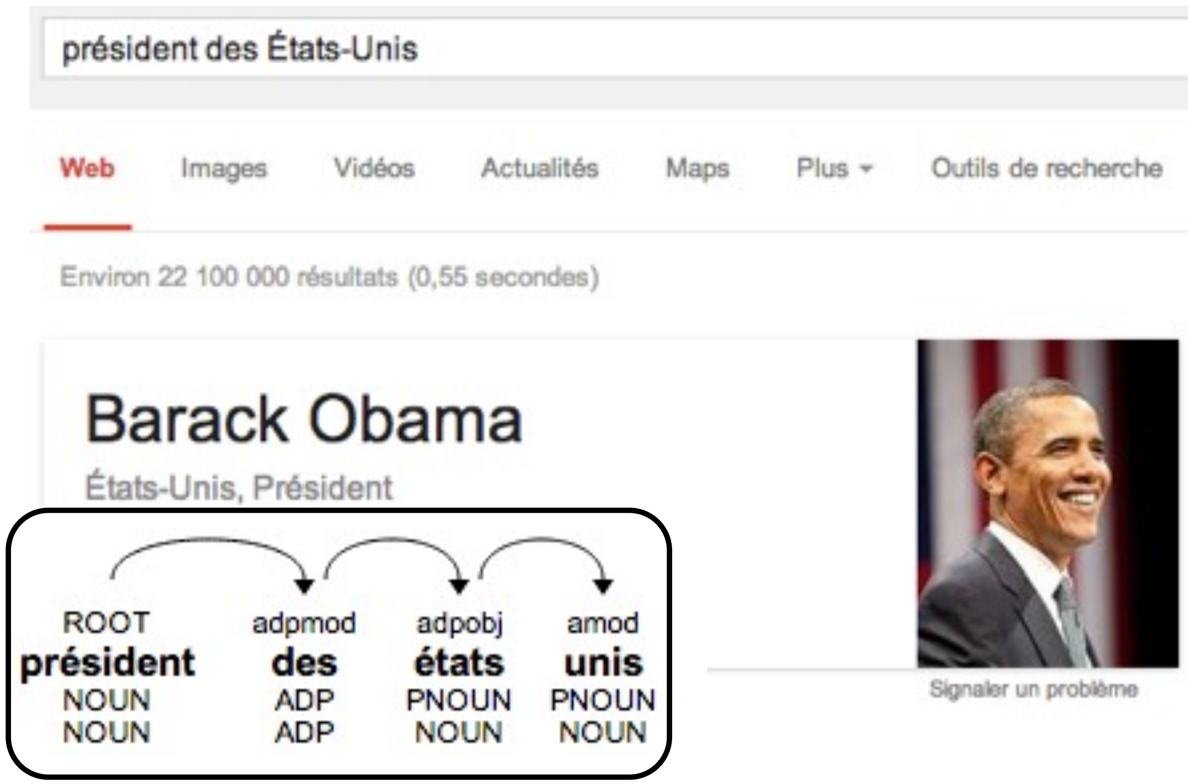


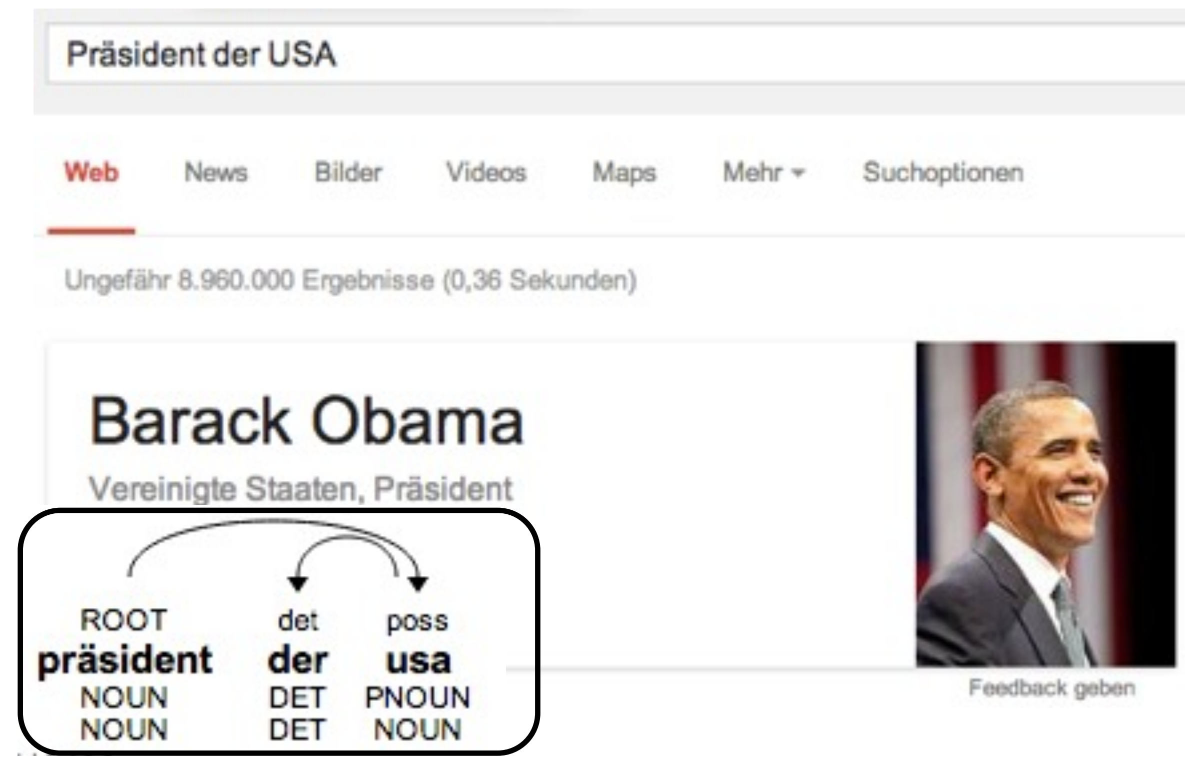


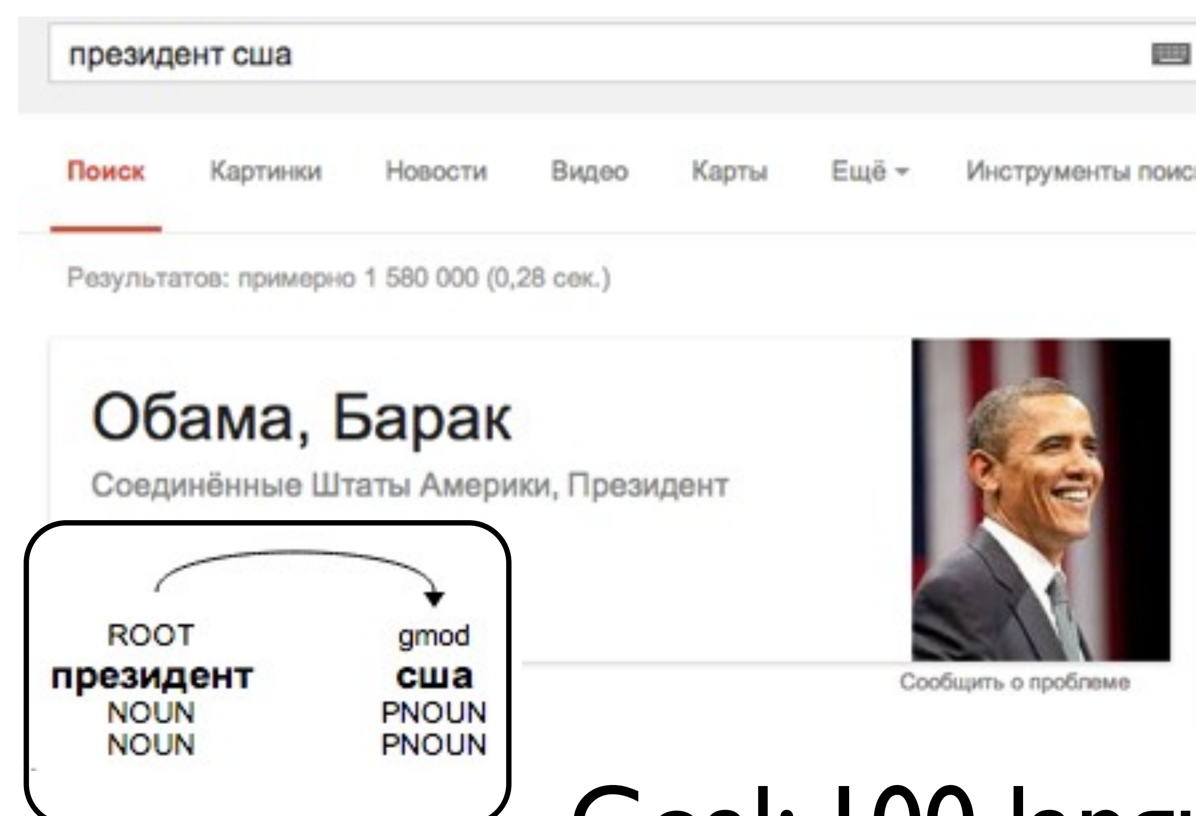


Goal: 100 languages

Natural Language Processing in Practice

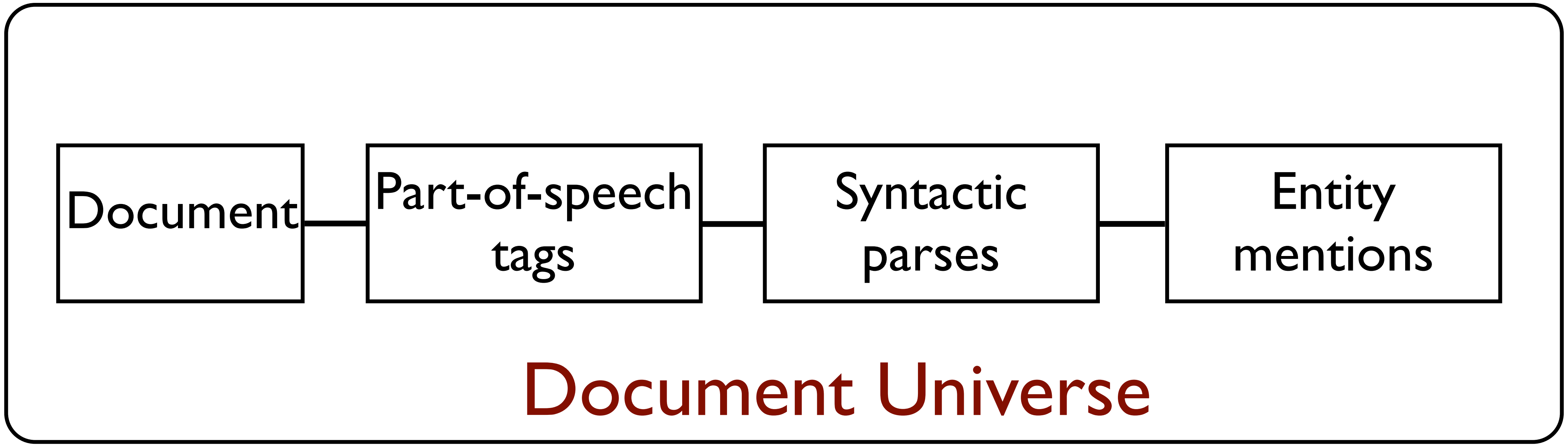




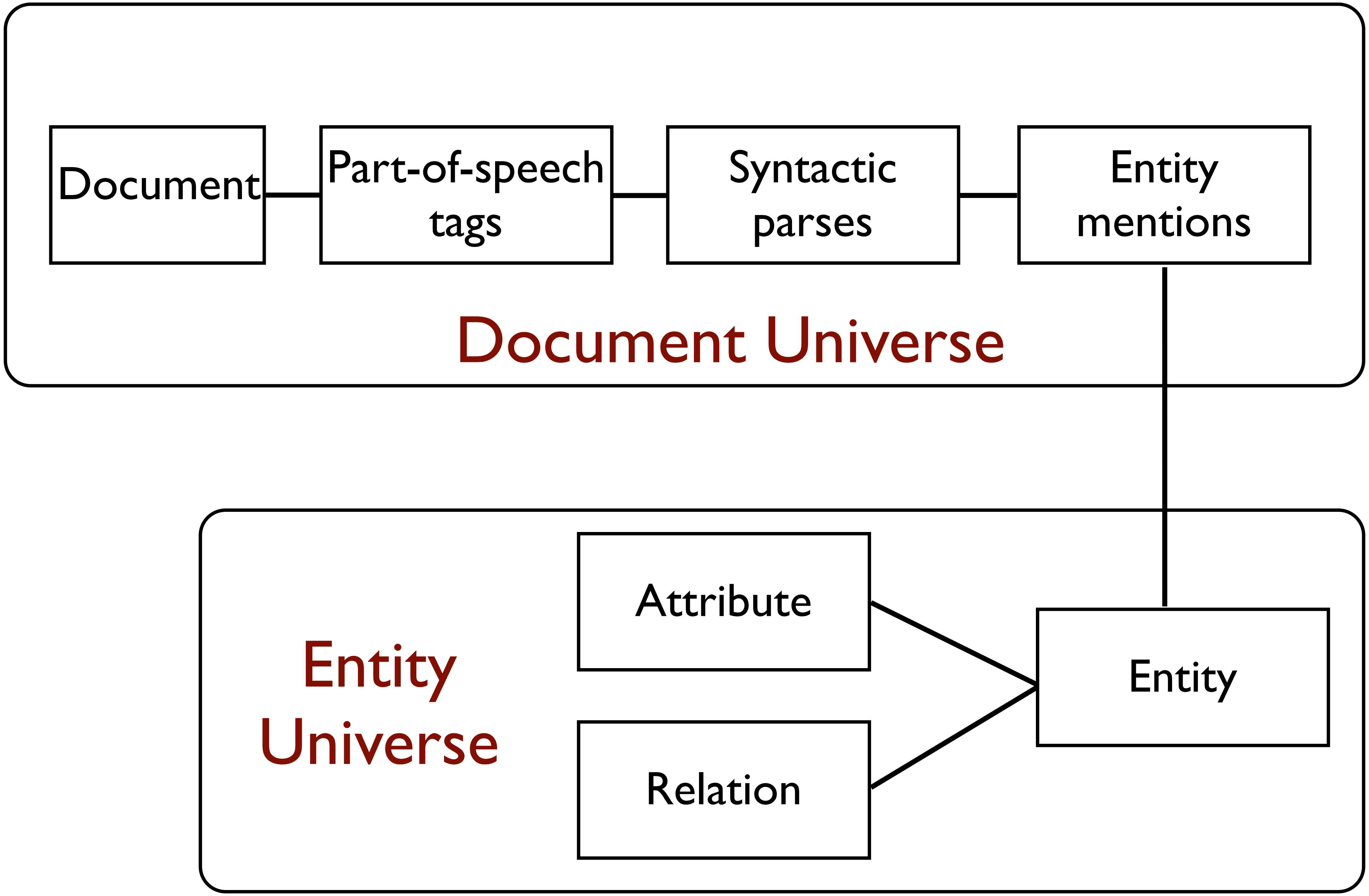


Goal: 100 languages

Natural Language Processing in Practice Google



Natural Language Processing in Practice

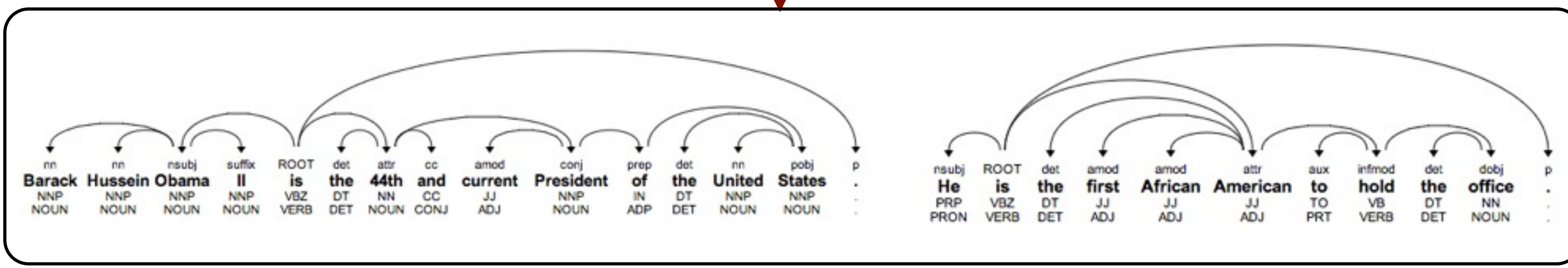


Natural Language Processing in Practice

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office.

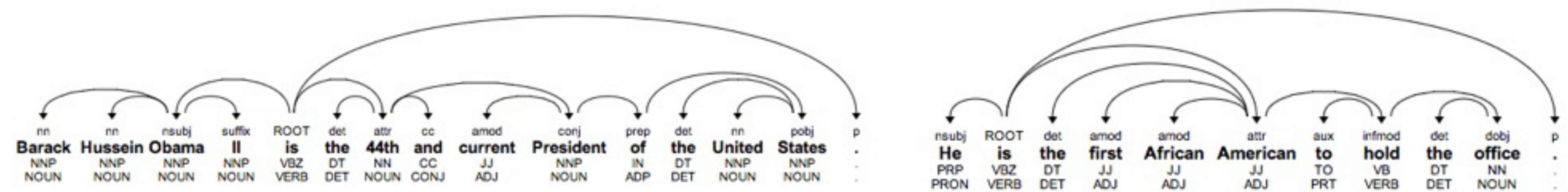
Natural Language Processing in Practice

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office.



Natural Language Processing in Practice Google

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office.



$_s$[Barack Hussein Obama II]$_1$ is the ⌞44th⌟ and current ⟨President⟩ $_2$ of the [United States]$_3$. $_s${He}$_1$ is the first [African American]$_4$ to hold the ⟨office⟩ $_5$.

Nat ... Practice Google

B ... he 44th and
c ... States. He is
t ... the office.

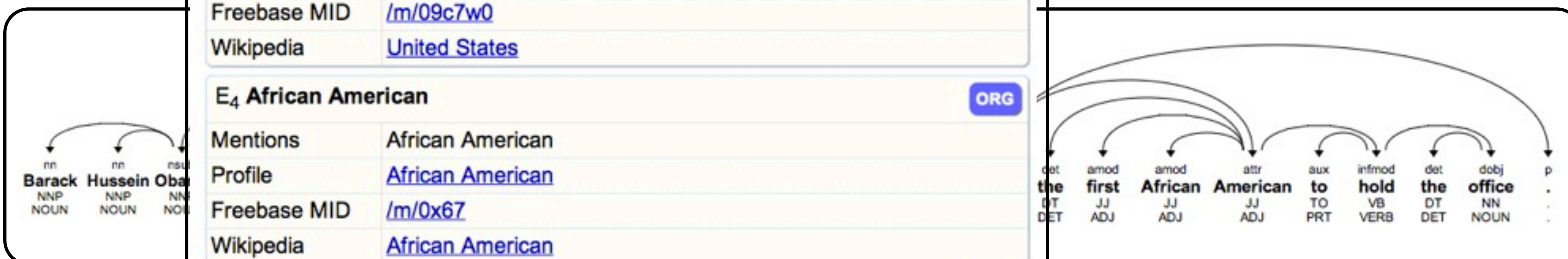
E_1 Barack Hussein Obama II	PER
Mentions	Barack Hussein Obama II, He
Gender	Male ♂
Profile	Barack Obama
Freebase MID	/m/02mjmr
Wikipedia	Barack Obama

E_2 President	PER
Mentions	President

E_3 United States	LOC
Mentions	United States
Profile	United States
Freebase MID	/m/09c7w0
Wikipedia	United States

E_4 African American	ORG
Mentions	African American
Profile	African American
Freebase MID	/m/0x67
Wikipedia	African American

E_5 office	ORG
Mentions	office



$_s$[Barack Hussein Obama II]$_1$ is the $_\llcorner$44th$_\lrcorner$ and current ⟨President⟩$_2$ of the [United States]$_3$. $_s${He}$_1$ is the first [African American]$_4$ to hold the ⟨office⟩$_5$.

NLP Pipeline

NLP Pipeline

- Build accurate natural language systems
  - Part-of-speech tagging
  - Parsing
  - Entity mention detection
  - Entity Linking
  - Relation Extraction
  - ...

NLP Pipeline

- Build accurate natural language systems
  - Part-of-speech tagging
  - Parsing
  - Entity mention detection
  - Entity Linking
  - Relation Extraction
  - ...
- Goal: Build such systems in ~100 languages

NLP Pipeline

Google

- Build accurate natural language systems
  - Part-of-speech tagging
  - Parsing
  - Entity mention detection
  - Entity Linking
  - Relation Extraction
  - ...

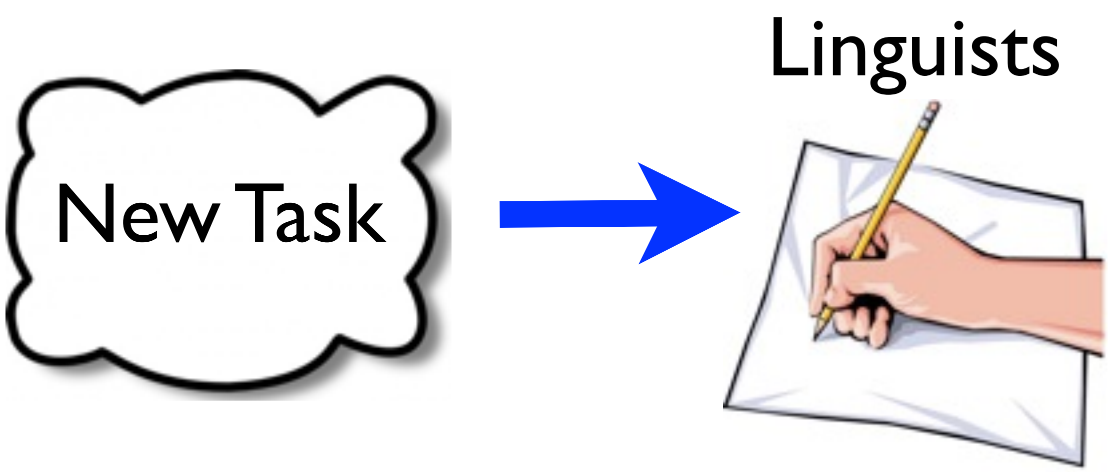
Syntax

- Goal: Build such systems in ~100 languages

Traditional Supervised NLP

New Task

Traditional Supervised NLP

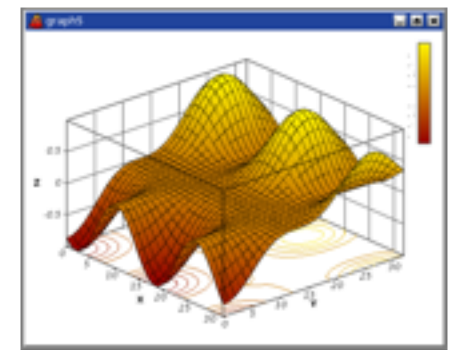


Traditional Supervised NLP

New Task

Linguists

Machine Learning

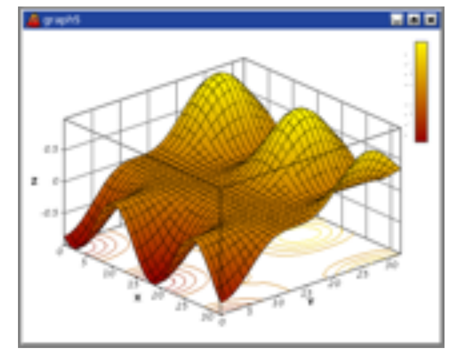


Traditional Supervised NLP

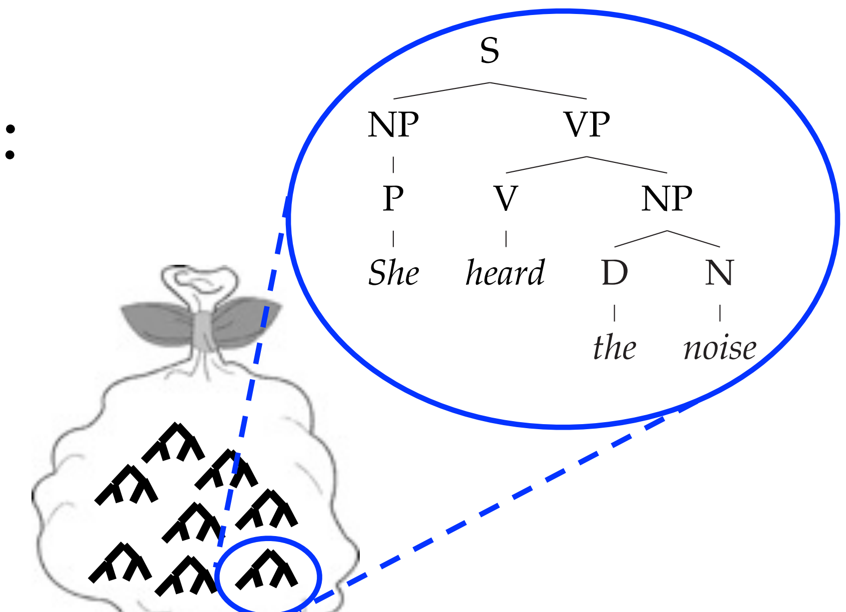
New Task

Linguists

Machine Learning

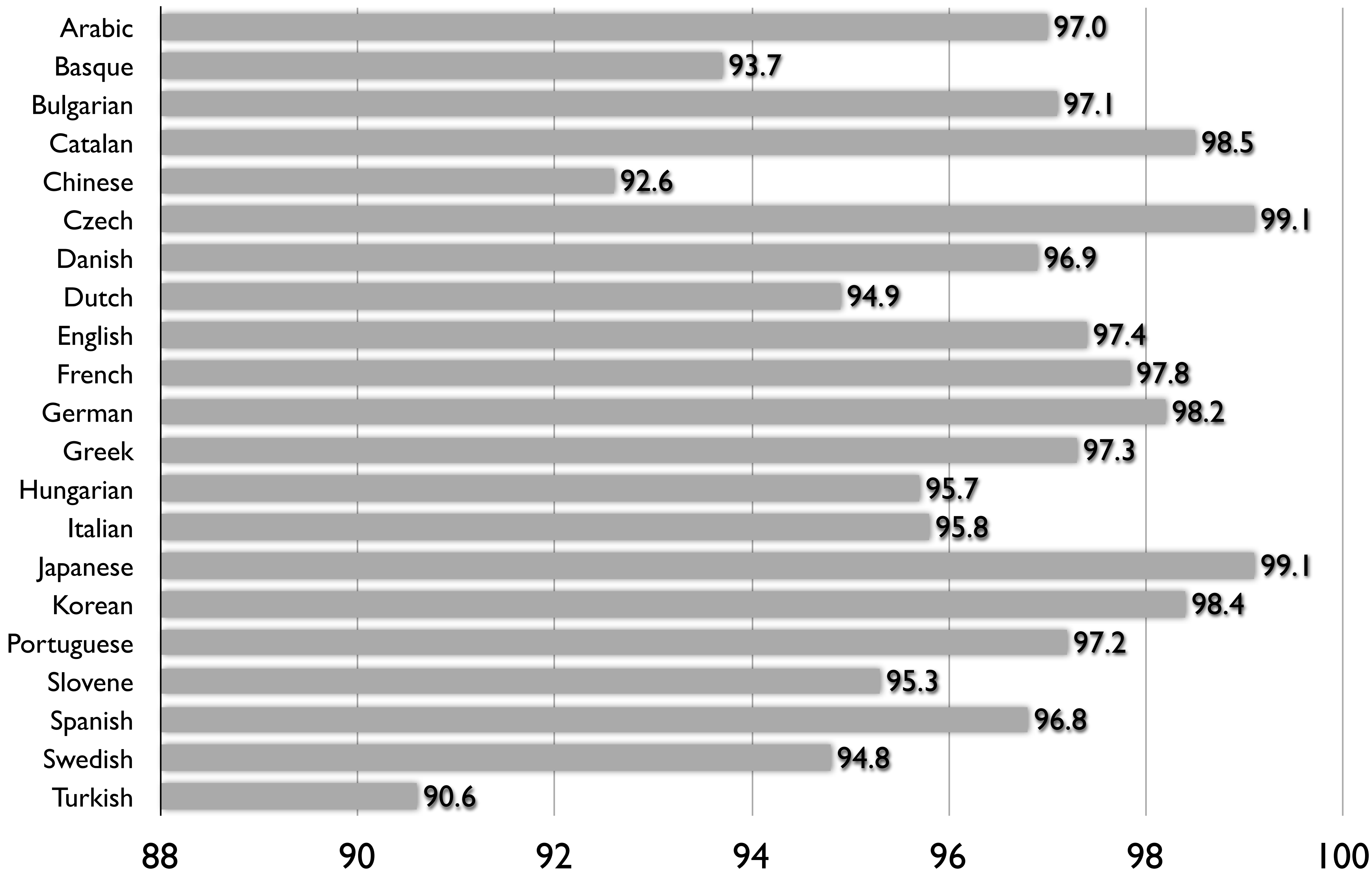


- Example: Syntactic analysis for English:
  - Not solved, but accuracies are high
  - 97% for parts-of-speech
  - 93% for parse trees

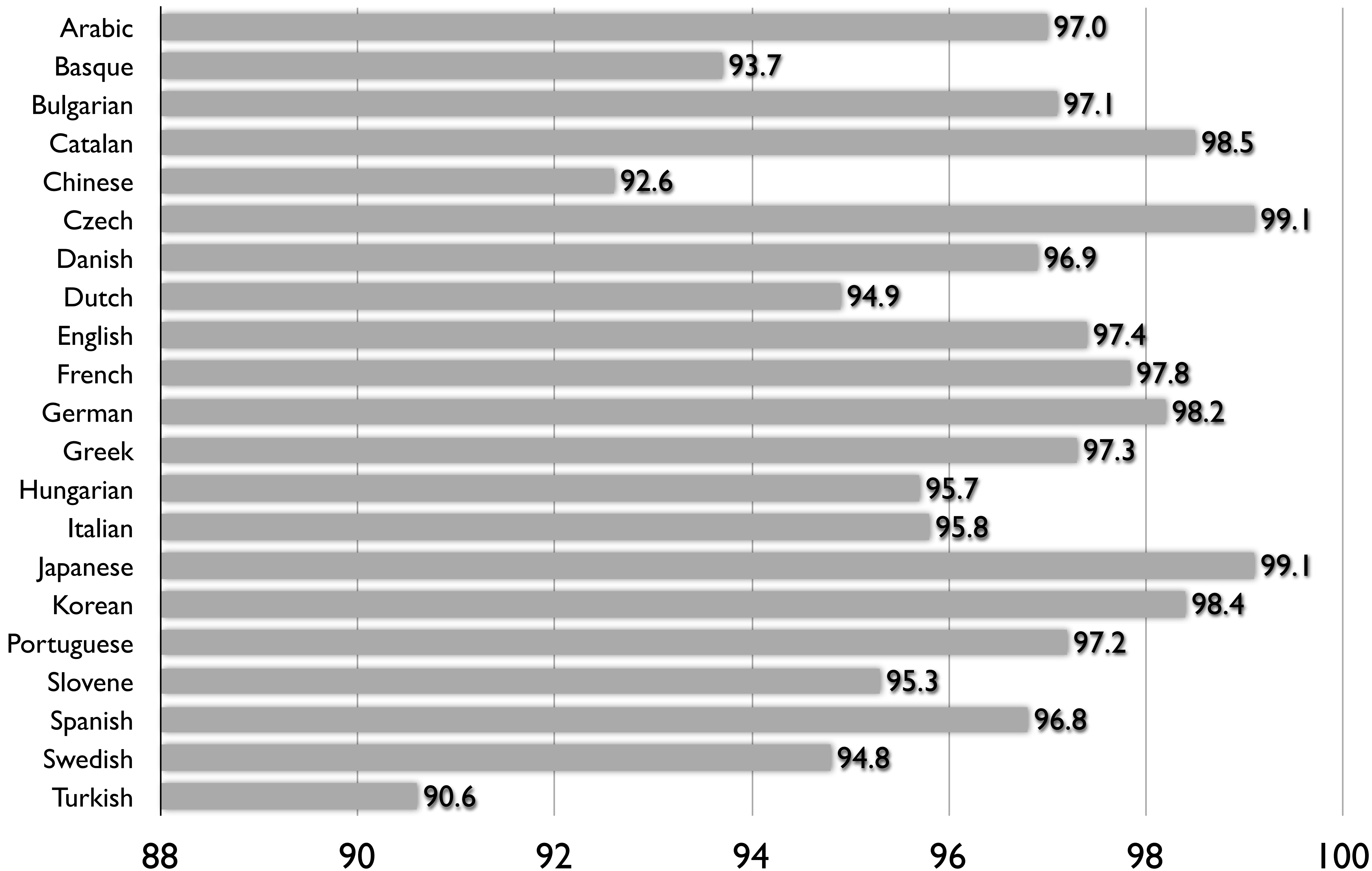


Treebank:
2 million words

Supervised Part-of-Speech Tagging



Supervised Part-of-Speech Tagging



Average accuracy is 96.2% (Brants, 2000)

Resource-Poor Languages

Google

Several major languages with little or no data

e.g	*Native speakers*
Punjabi	109 million
Vietnamese	69 million
Oriya	32 million
Indonesian-Malay	37 million
Azerbaijani	20 million
Haitian	7.7 million

Resource-Poor Languages

Several major languages with little or no data

e.g	*Native speakers*
Punjabi	109 million
Vietnamese	69 million
Oriya	32 million
Indonesian-Malay	37 million
Azerbaijani	20 million
Haitian	7.7 million

However, lots of translations to English and unlabeled data

Resource-Poor Languages

Several major languages with little or no data

e.g	*Native speakers*
Punjabi	109 million
Vietnamese	69 million
Oriya	32 million
Indonesian-Malay	37 million
Azerbaijani	20 million
Haitian	7.7 million

However, lots of translations to English and unlabeled data

Basic tools like POS taggers essential for language technologies

Talk Outline

Talk Outline

- Cross-lingual learning for part-of-speech tagging
  - Use of translated data from English to target languages
  - Crowdsourced dictionaries

Talk Outline

- Cross-lingual learning for part-of-speech tagging
  - Use of translated data from English to target languages
  - Crowdsourced dictionaries
- A word about cross-lingual learning for syntactic parsing

Cross-Lingual Part-of-Speech Tagging

Learning Part-of-Speech Taggers with Projected Dictionaries

Cross-Lingual Part-of-Speech Tagging

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

Cross-Lingual Part-of-Speech Tagging

Learning Part-of-Speech Taggers with Projected Dictionaries

Das and Petrov (ACL 2011)

Incorporating Hard Constraints from Translated sentences and Dictionaries

Täckström, Das, Petrov, McDonald and Nivre (TACL 2013)

Generalizing above with *Posterior Regularization*

Ganchev and Das (EMNLP 2013)

Cross-Lingual Part-of-Speech Tagging

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

Part-of-Speech Tagging

The food at Google is good .

Part-of-Speech Tagging

The	food	at	Google	is	good	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

(Nearly) Universal Part-of-Speech Tags Google

VERB	DET
NOUN	CONJ
PRON	NUM
ADJ	PRT
ADV	.
ADP	X

(Petrov, Das and McDonald, 2012)

(Nearly) Universal Part-of-Speech Tags Google

The food at Google is good .

A comida no Google é boa .

गूगल में खाना अच्छा है .

(Nearly) Universal Part-of-Speech Tags

The	food	at	Google	is	good	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

A comida no Google é boa .

गूगल में खाना अच्छा है .

(Nearly) Universal Part-of-Speech Tags Google

The	food	at	Google	is	good	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

A	comida	no	Google	é	boa	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

गूगल	में	खाना	अच्छा	है	.

(Nearly) Universal Part-of-Speech Tags Google

The	food	at	Google	is	good	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

A	comida	no	Google	é	boa	.
DET	NOUN	ADP	NOUN	VERB	ADJ	.

गूगल	में	खाना	अच्छा	है	.
NOUN	ADP	NOUN	ADJ	VERB	.

State of the Art in *Unsupervised* Part-of-Speech Tagging

Unsupervised Part-of-Speech Tagging

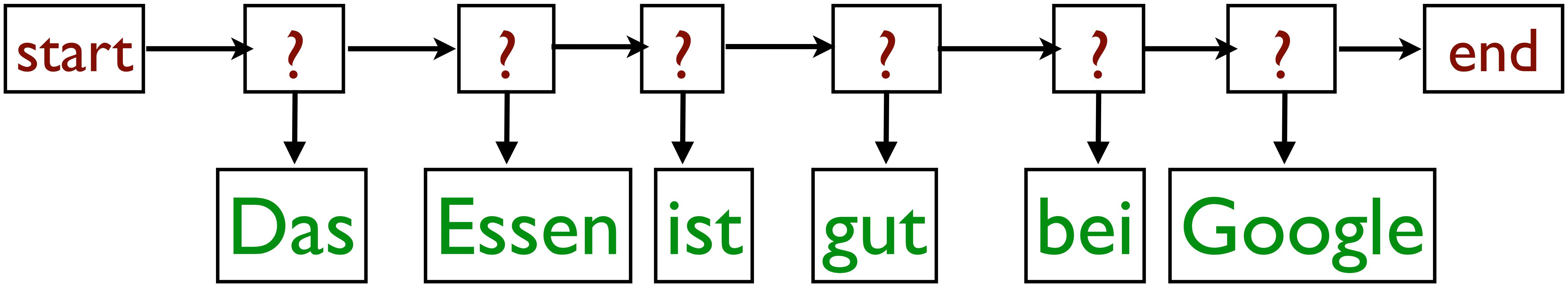
Hidden Markov Model estimated with the Expectation-Maximization algorithm

Das Essen ist gut bei Google

Merialdo (1994)

Unsupervised Part-of-Speech Tagging

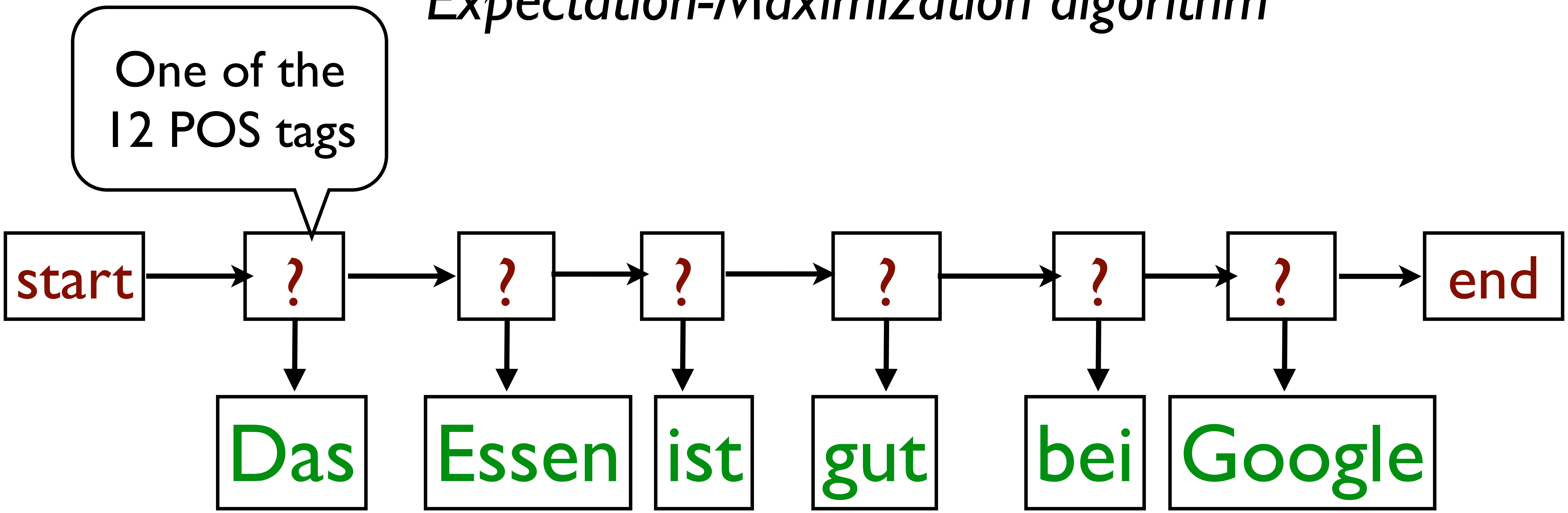
Hidden Markov Model estimated with the Expectation-Maximization algorithm



Merialdo (1994)

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm



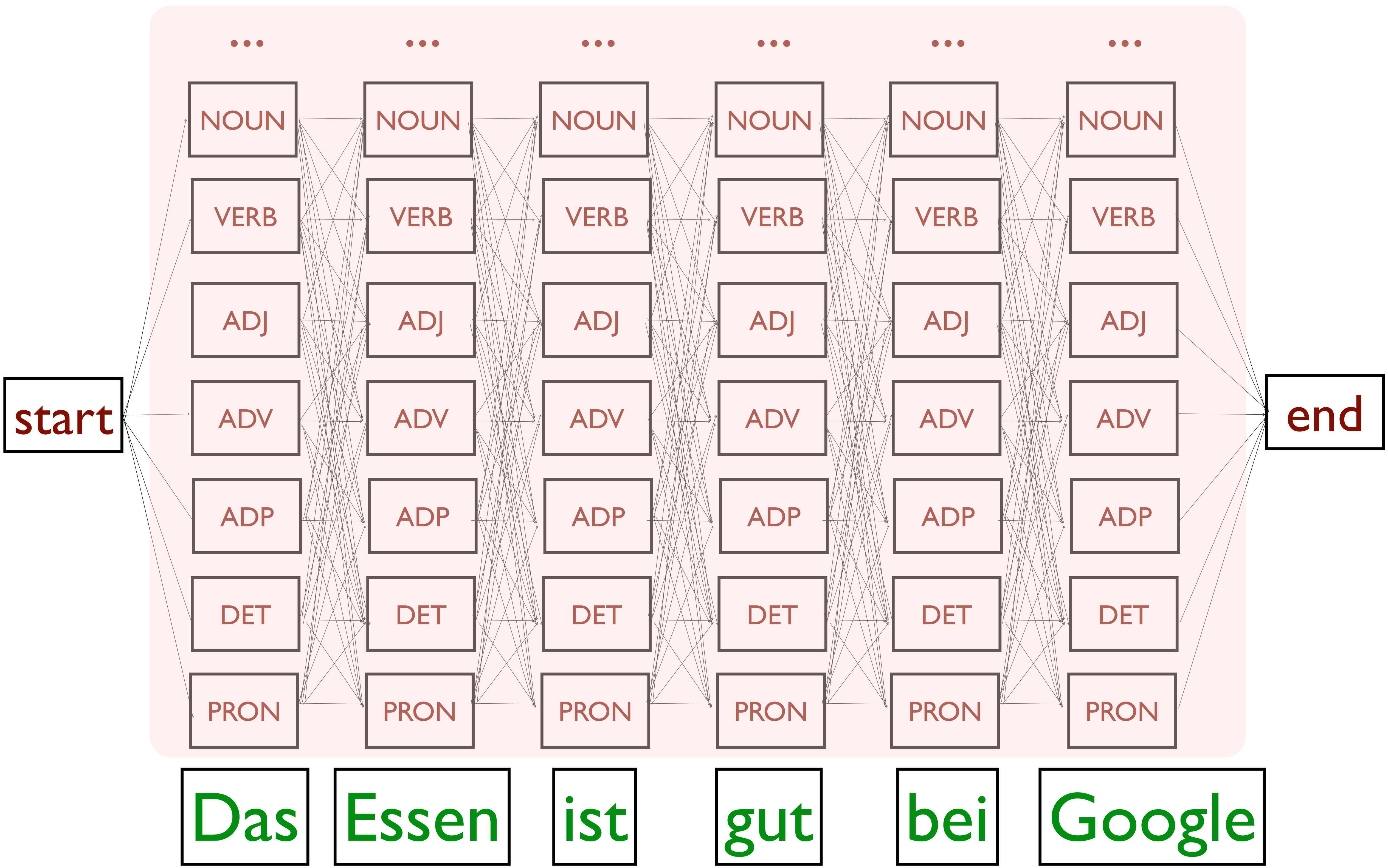
Merialdo (1994)

Unsupervised Part-of-Speech Tagging

Das Essen ist gut bei Google

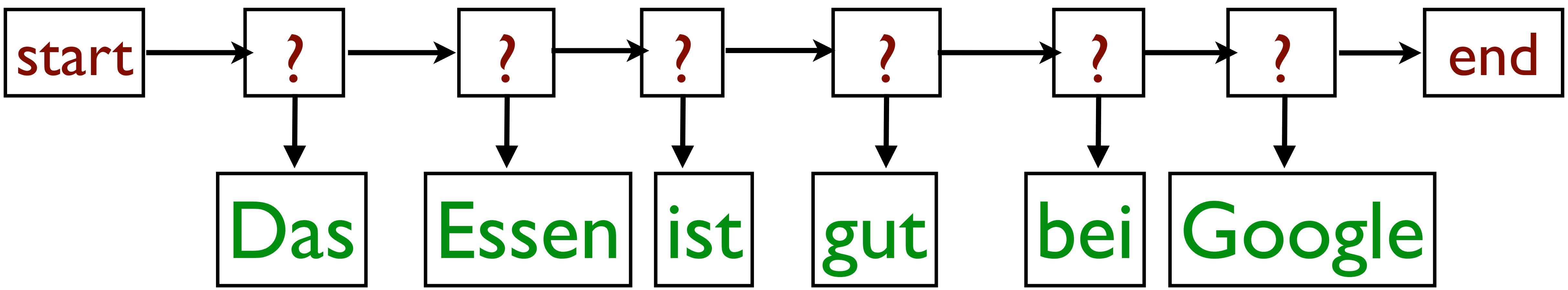
Unsupervised Part-of-Speech Tagging

HMM lattice



Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm

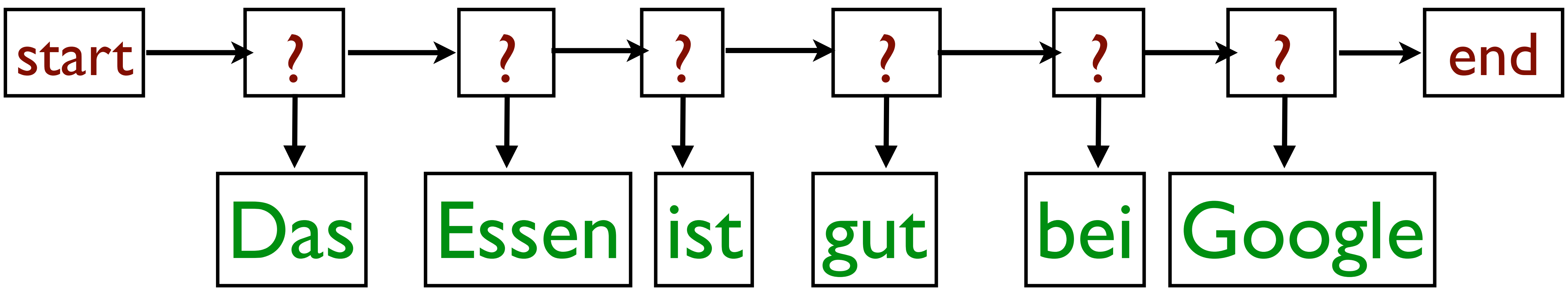


x : word sequence

y : tag sequence

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm



$\boldsymbol{x}$: word sequence

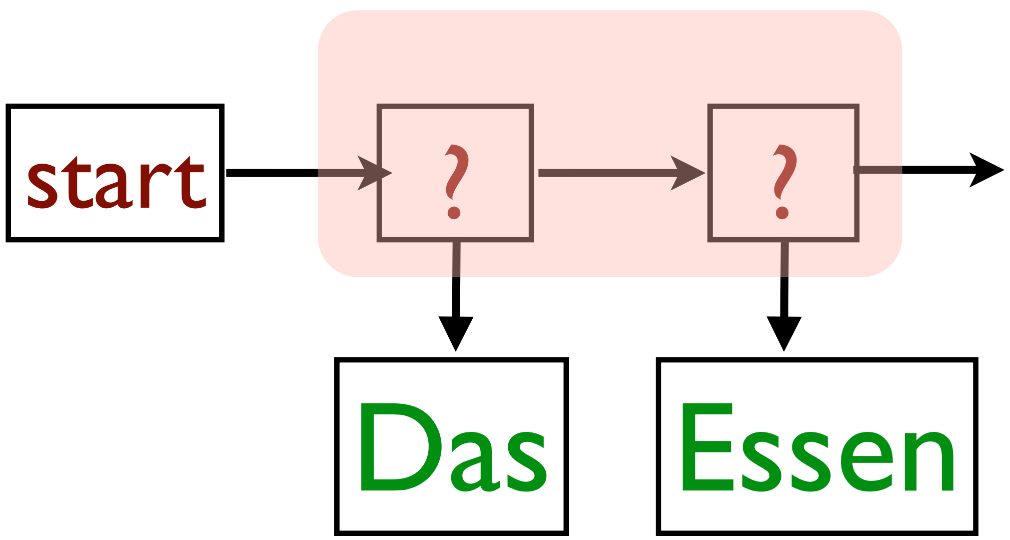
$\boldsymbol{y}$: tag sequence

Model

$$p(\boldsymbol{x}, \boldsymbol{y})$$

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm



transition multinomials

$$p(y_i | y_{i-1})$$

$\boldsymbol{x}$: word sequence

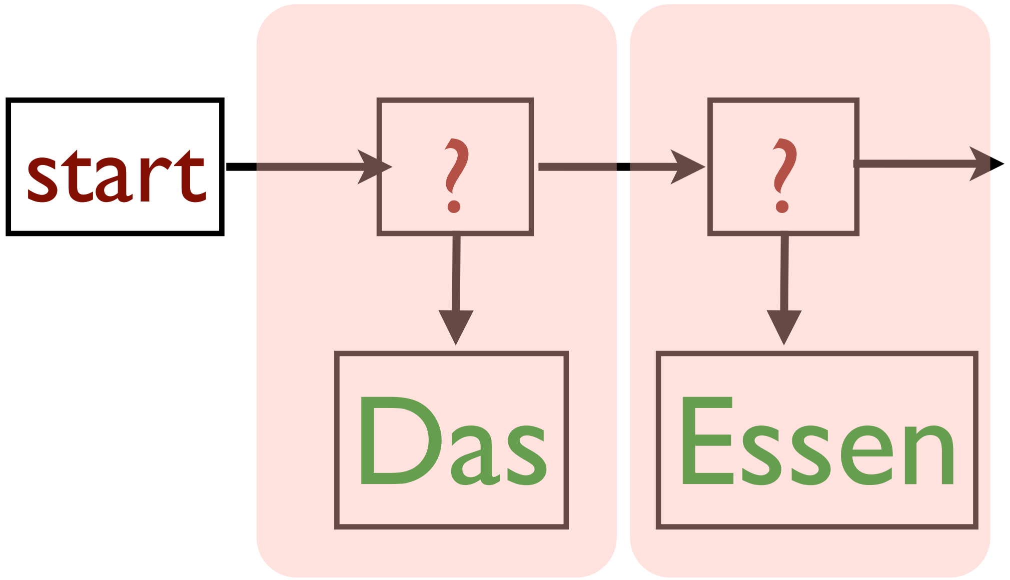
$\boldsymbol{y}$: tag sequence

Model

$$p(\boldsymbol{x}, \boldsymbol{y})$$

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm



emission multinomials

$$p(x_i | y_i)$$

$\boldsymbol{x}$: word sequence

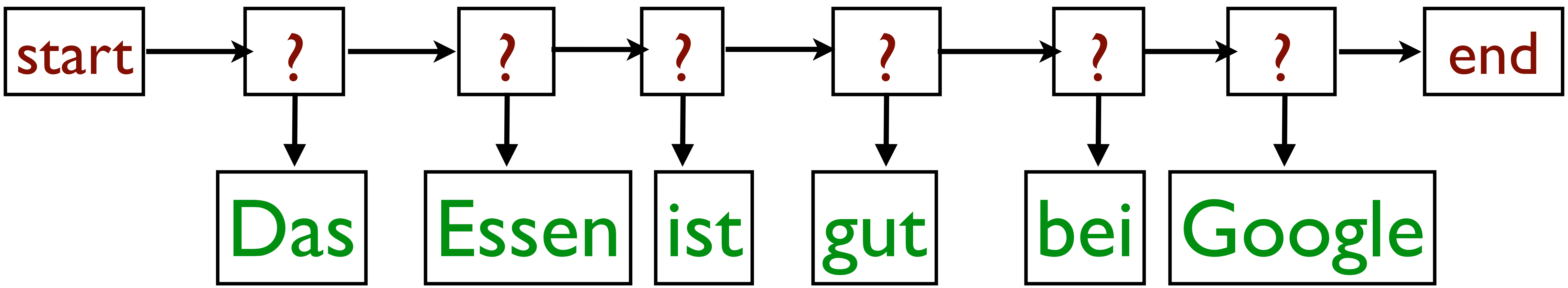
$\boldsymbol{y}$: tag sequence

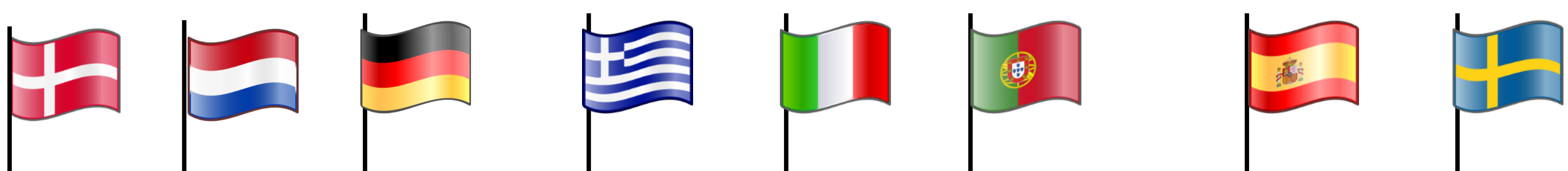
Model

$$p(\boldsymbol{x}, \boldsymbol{y})$$

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm



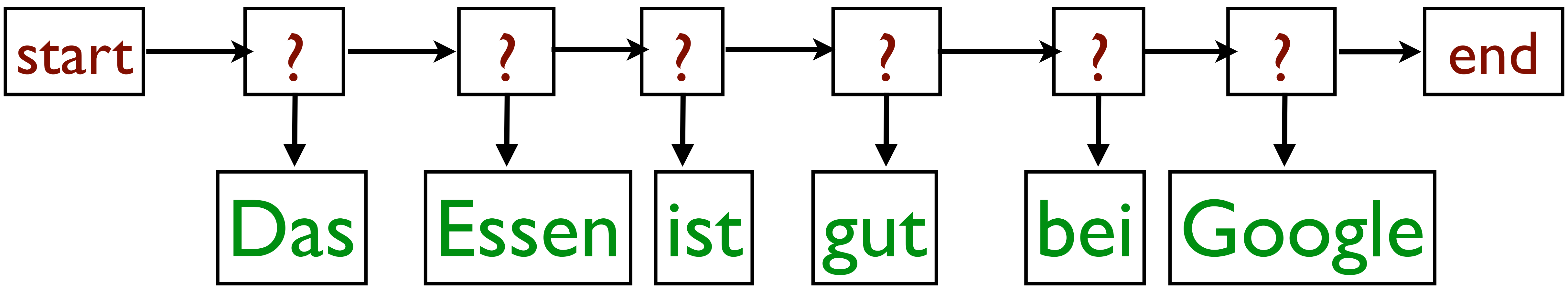


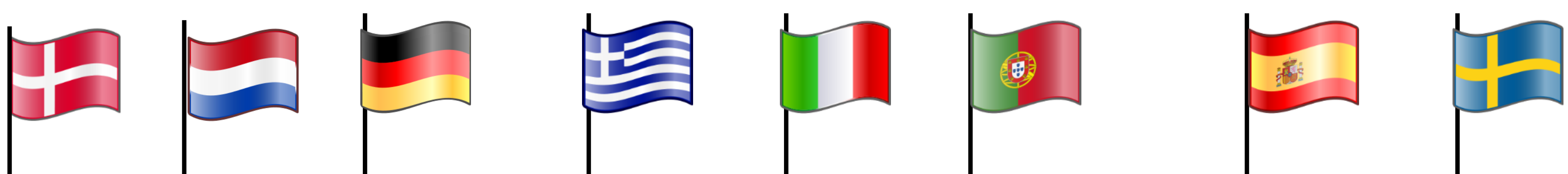
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7

Johnson (2007)

Unsupervised Part-of-Speech Tagging

Hidden Markov Model estimated with the Expectation-Maximization algorithm





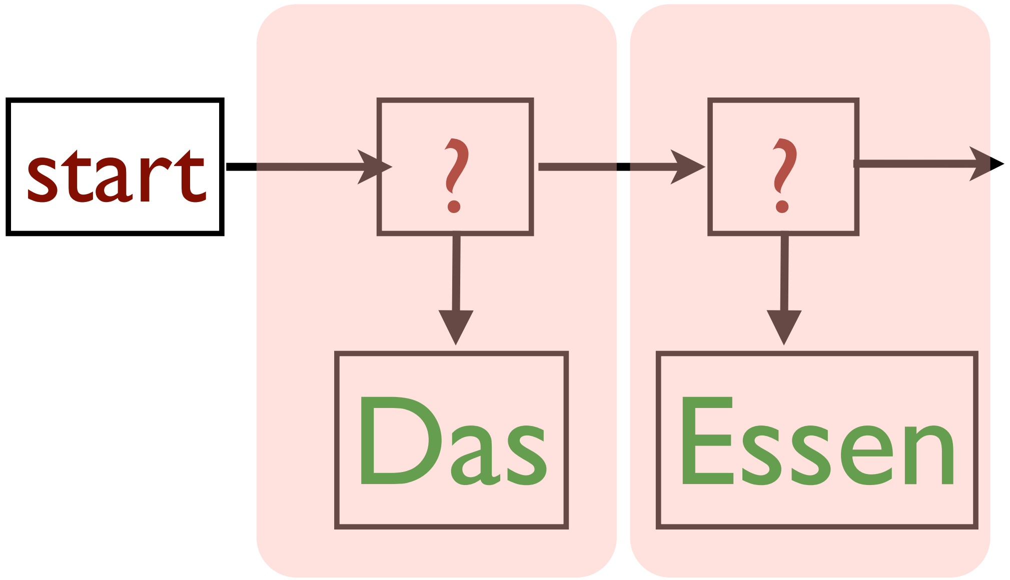
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7

Johnson (2007)

Poor average result

Unsupervised Part-of-Speech Tagging

HMM with locally-normalized log-linear models



emission multinomials

$$p(x_i|y_i) \propto \exp \boldsymbol{\theta} \cdot f(x_i, y_i)$$

Berg-Kirkpatrick et al. (2010)

$\boldsymbol{x}$: word sequence

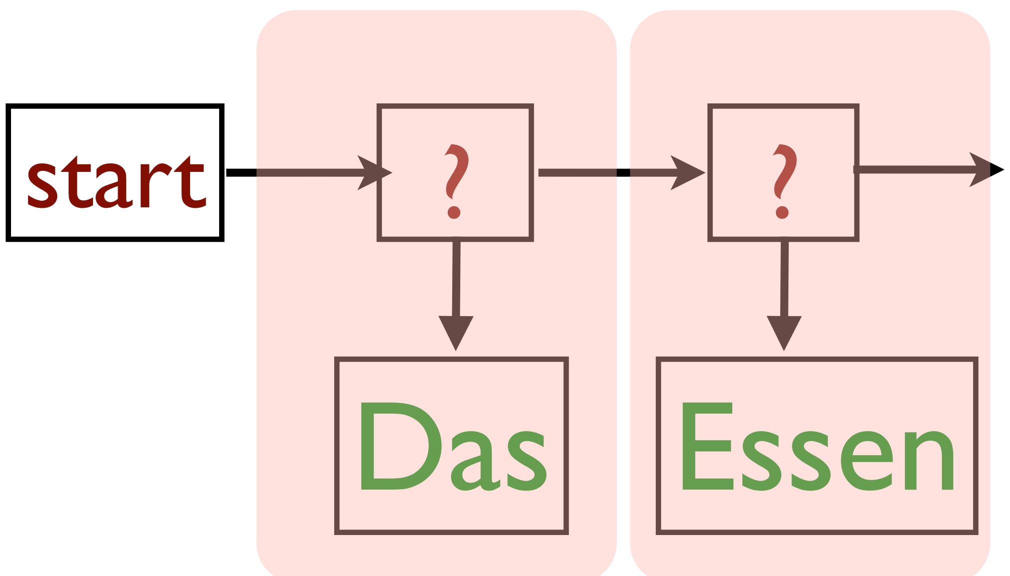
$\boldsymbol{y}$: tag sequence

Model

$$p(\boldsymbol{x}, \boldsymbol{y})$$

Unsupervised Part-of-Speech Tagging

HMM with locally-normalized log-linear models



emission multinomials

$$p(x_i|y_i) \propto \exp \boldsymbol{\theta} \cdot \boxed{f}(x_i, y_i)$$

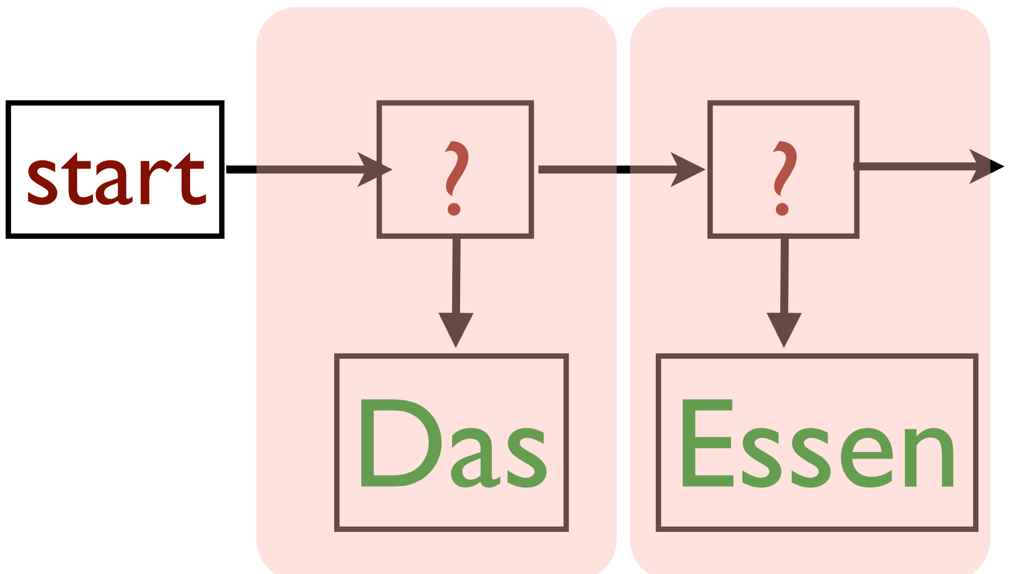
suffix
hyphen
capitalization
numbers

x : word sequence

y : tag sequence

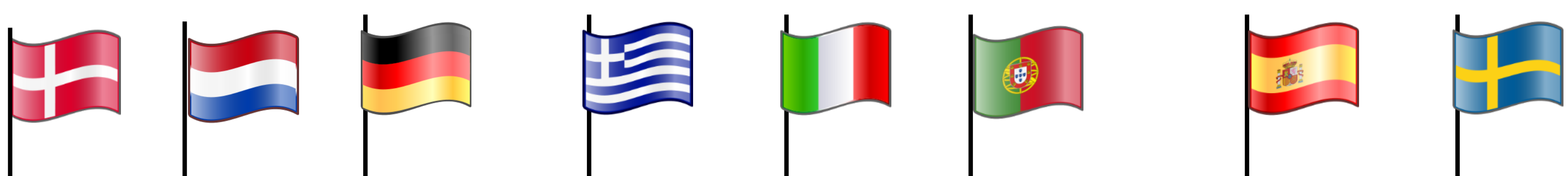
Unsupervised Part-of-Speech Tagging

HMM with locally-normalized log-linear models



emission multinomials

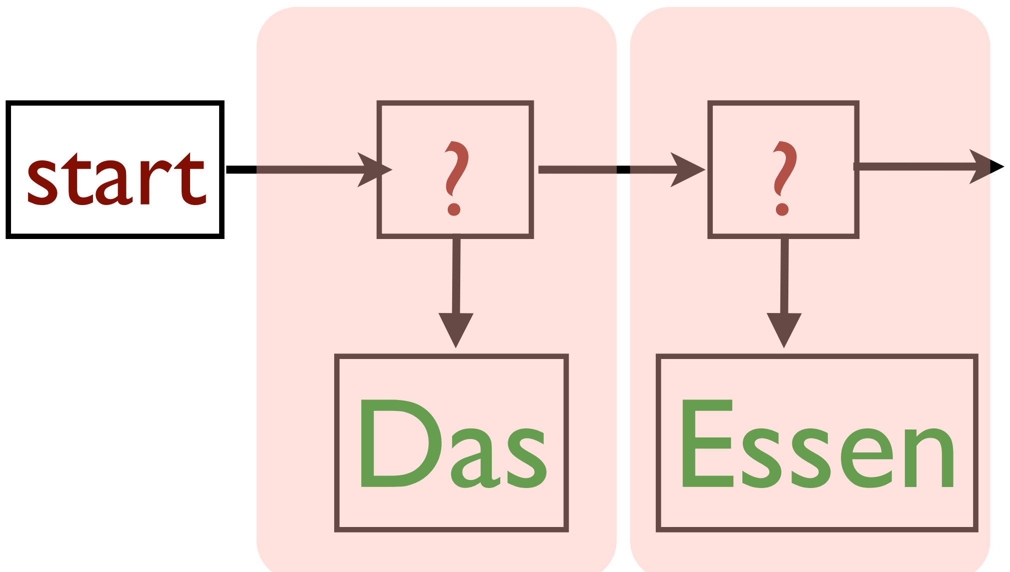
$$p(x_i|y_i) \propto \exp \boldsymbol{\theta} \cdot f(x_i, y_i)$$



	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	**69.1**	**65.1**	**81.3**	**71.8**	**68.1**	**78.4**	**80.2**	**70.1**	**73.0**

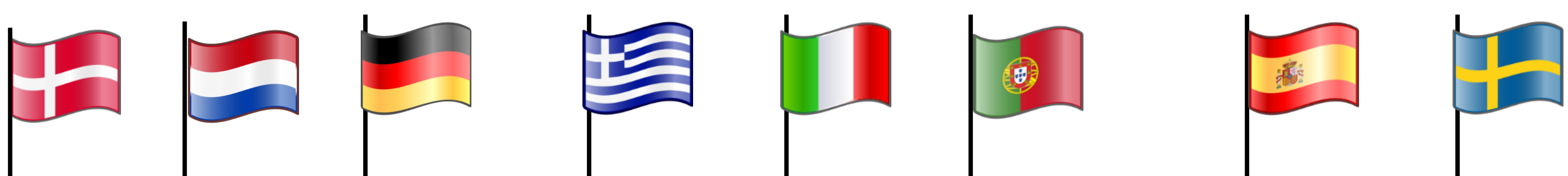
Unsupervised Part-of-Speech Tagging

HMM with locally-normalized log-linear models



emission multinomials

$$p(x_i|y_i) \propto \exp \boldsymbol{\theta} \cdot f(x_i, y_i)$$

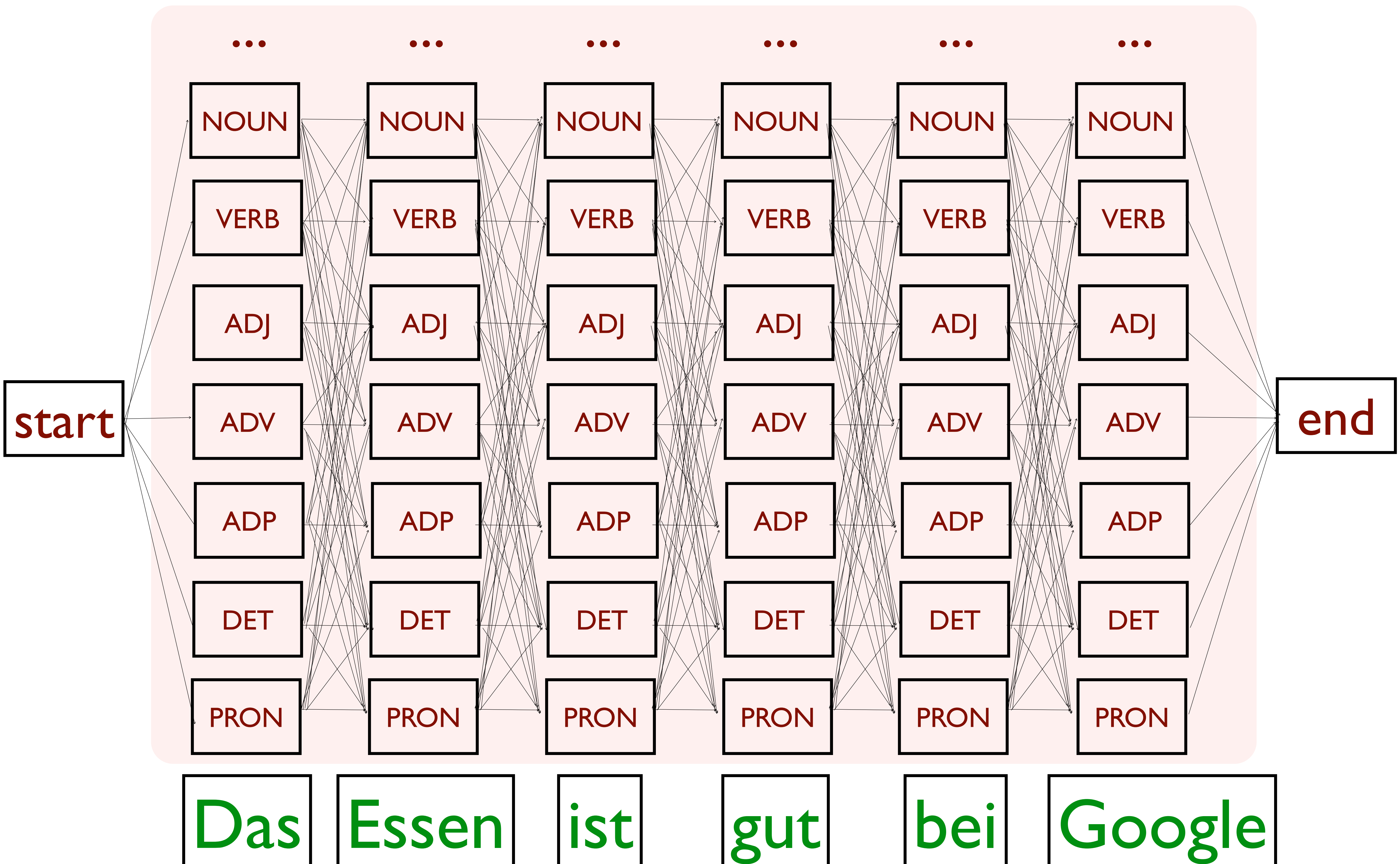


	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	**69.1**	**65.1**	**81.3**	**71.8**	**68.1**	**78.4**	**80.2**	**70.1**	**73.0**

Improvements across all languages

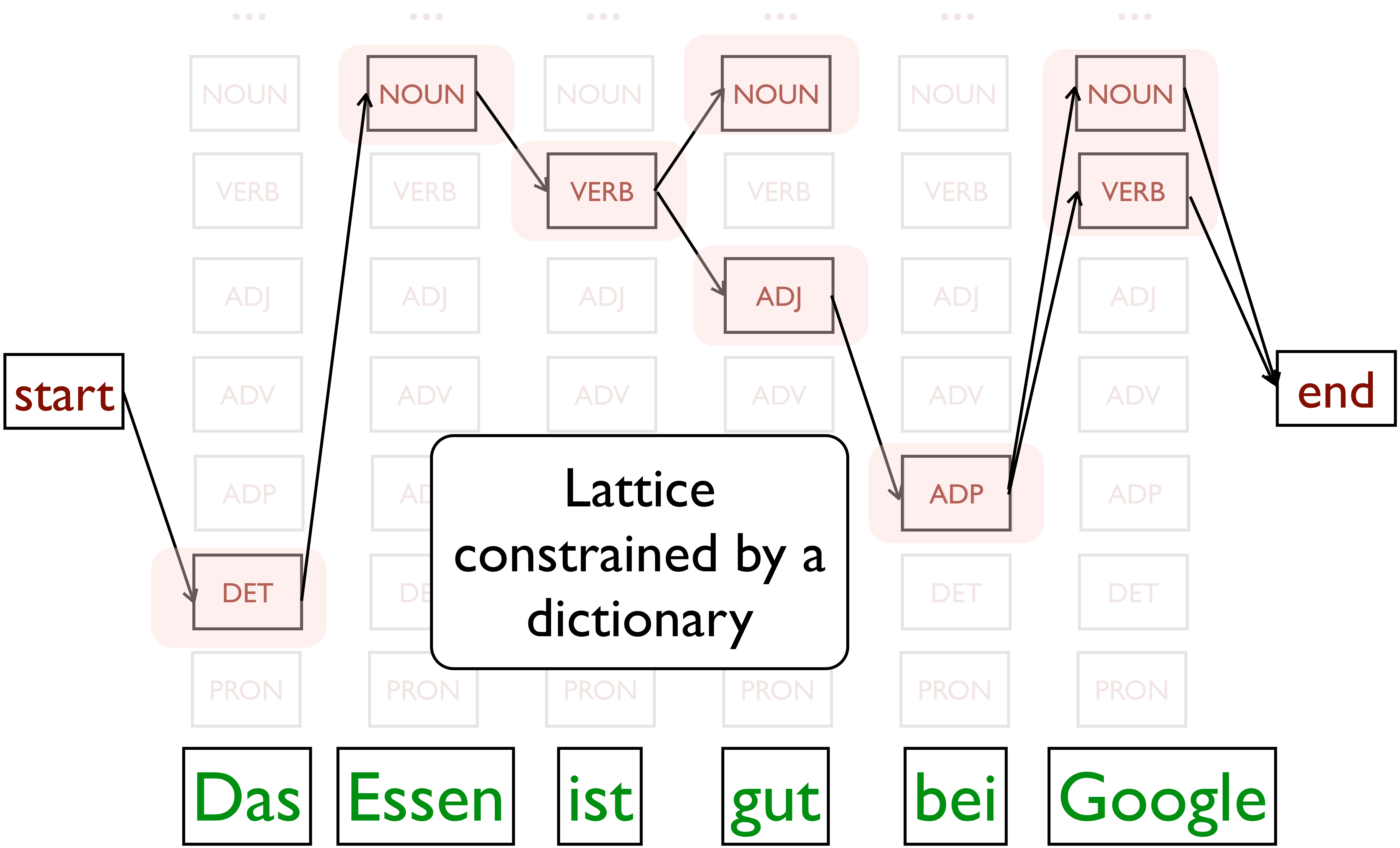
Unsupervised POS Tagging with **Dictionaries**

HMM with locally-normalized log-linear models



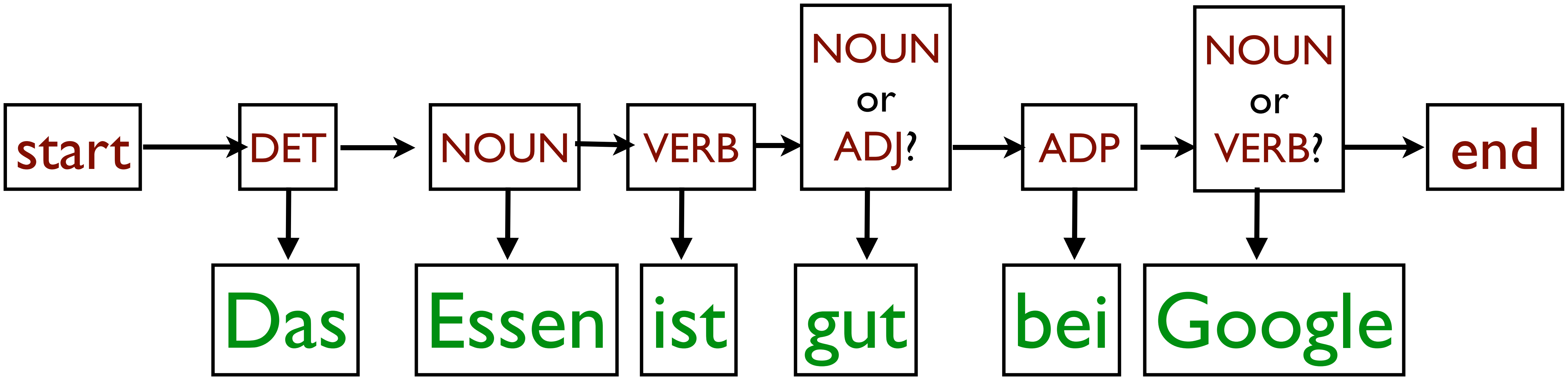
Unsupervised POS Tagging with **Dictionaries**

HMM with locally-normalized log-linear models



Unsupervised POS Tagging with **Dictionaries**

HMM with locally-normalized log-linear models



	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
w/ gold dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*

Can we construct dictionaries
for new languages easily?

Can we construct dictionaries
for new languages easily?

Ideas:

- Use supervised systems in resource-rich languages
- Use translations or **parallel** data with a target language
- Construct noisy, **projected** tag lexicons

Cross-Lingual Projection

Das Essen ist gut bei Google

Cross-Lingual Projection

The food at Google is good

Das Essen ist gut bei Google

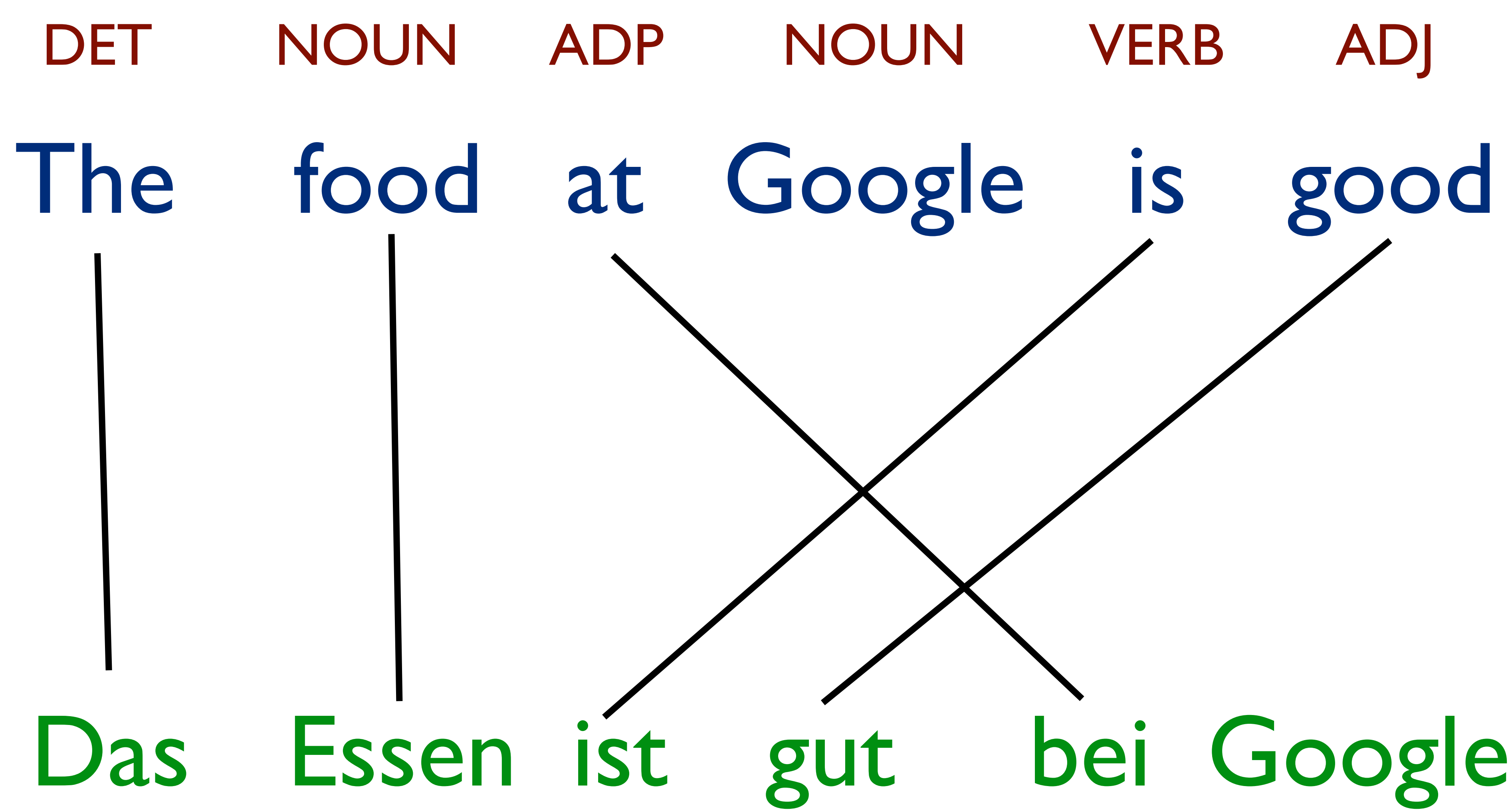
Cross-Lingual Projection

automatic labels from supervised tagger, 97% accuracy

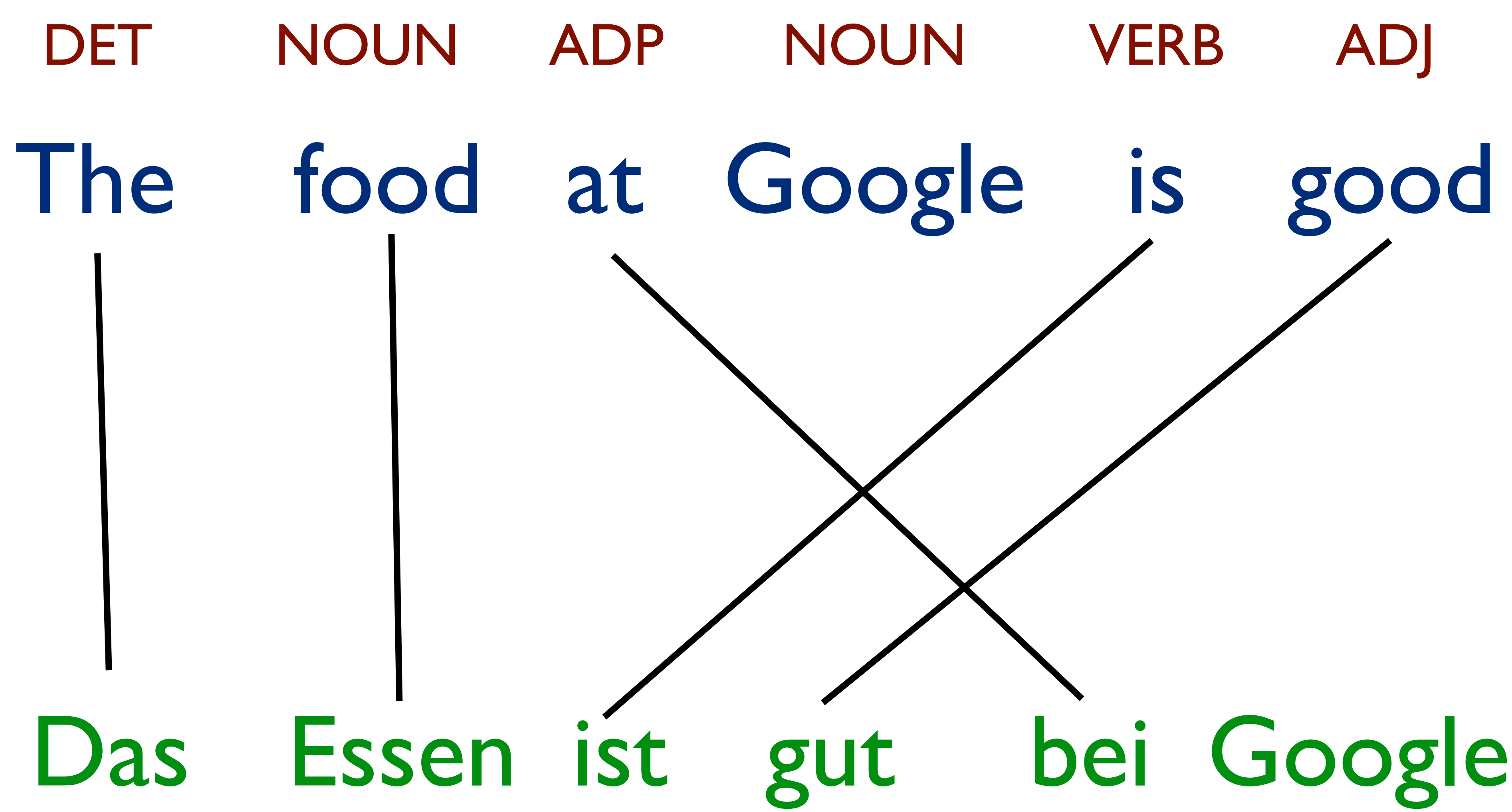
DET	NOUN	ADP	NOUN	VERB	ADJ
The	food	at	Google	is	good

Das Essen ist gut bei Google

Cross-Lingual Projection



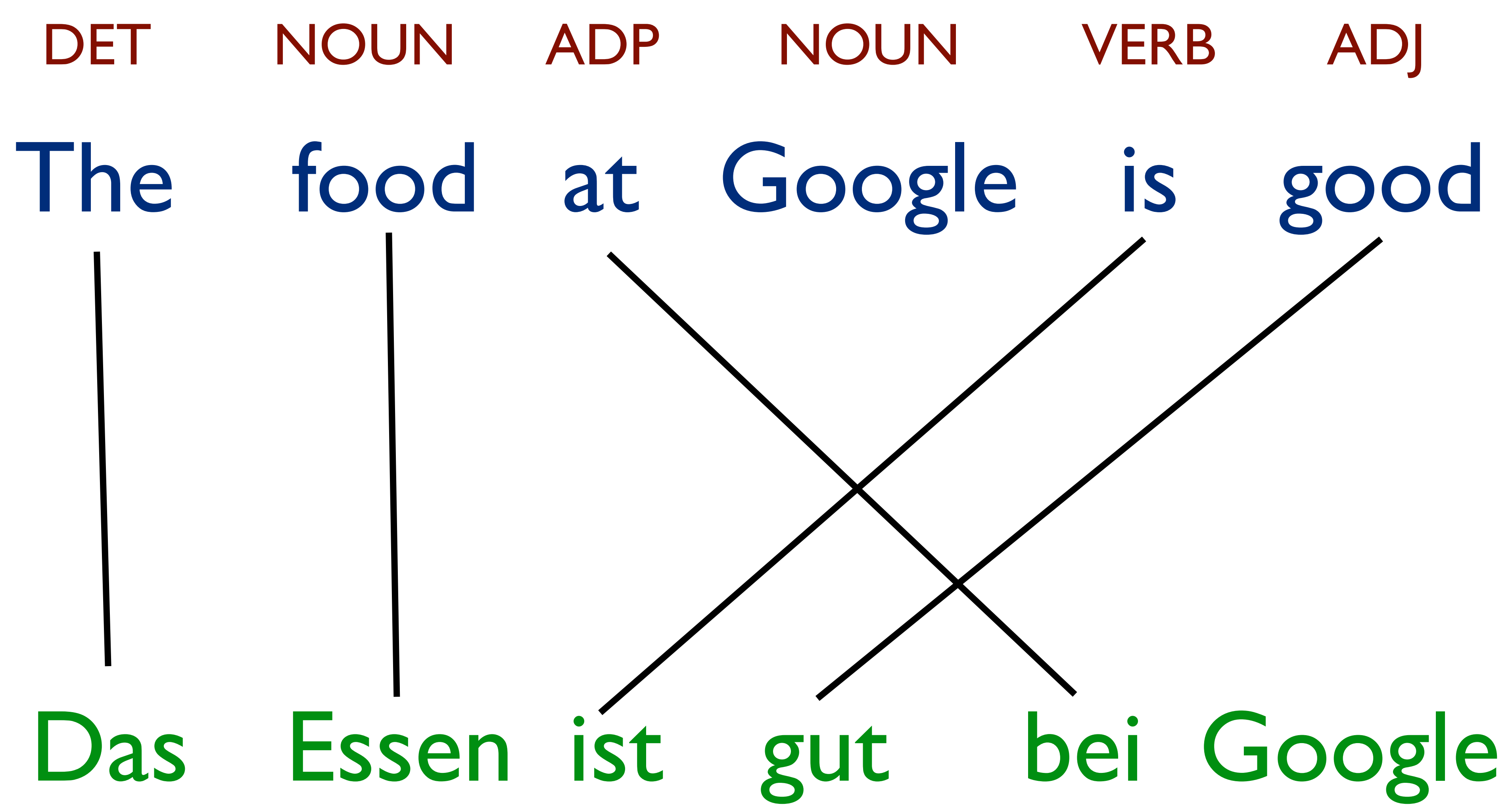
Cross-Lingual Projection



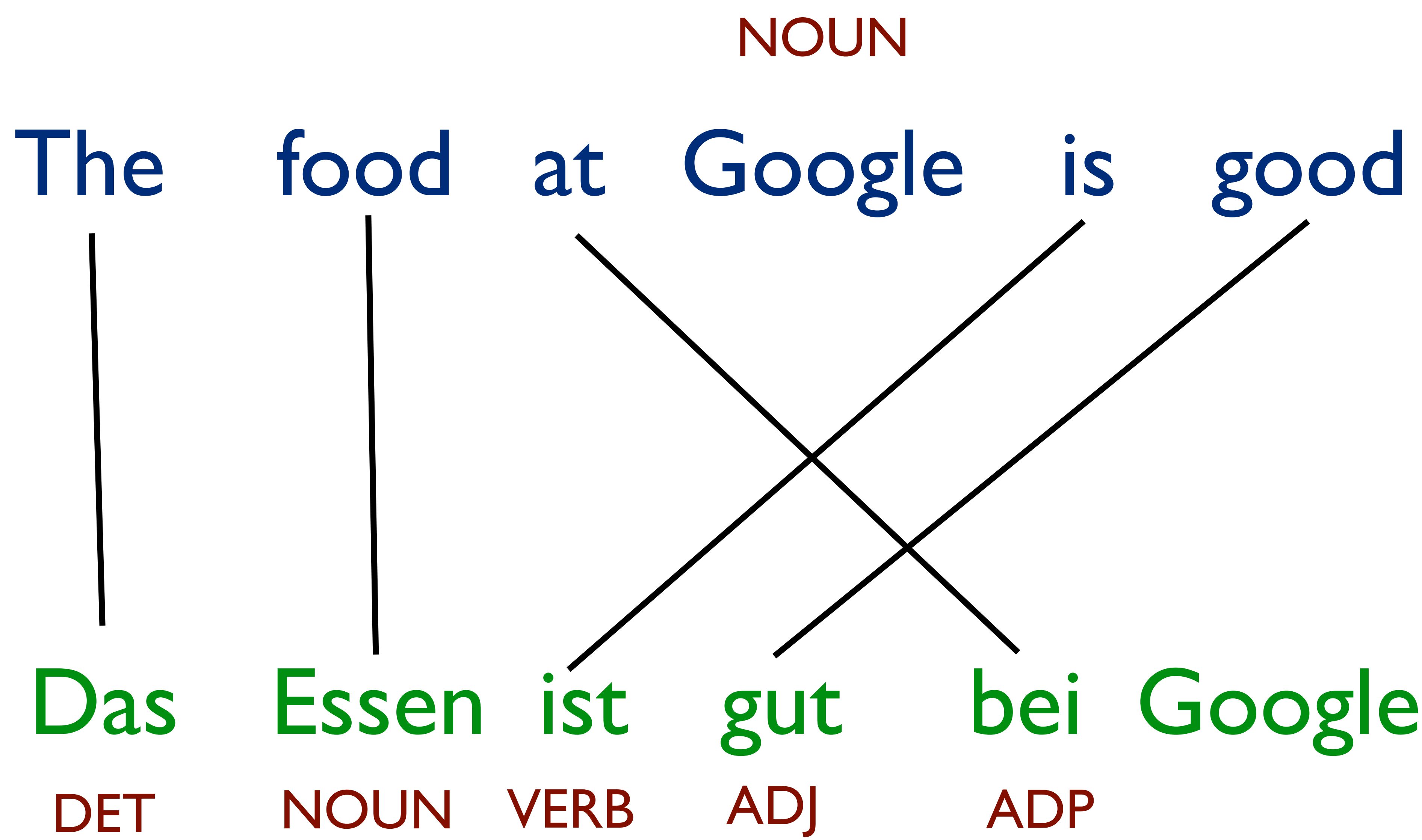
Automatic unsupervised alignments from translation data

(available for ~100 languages)

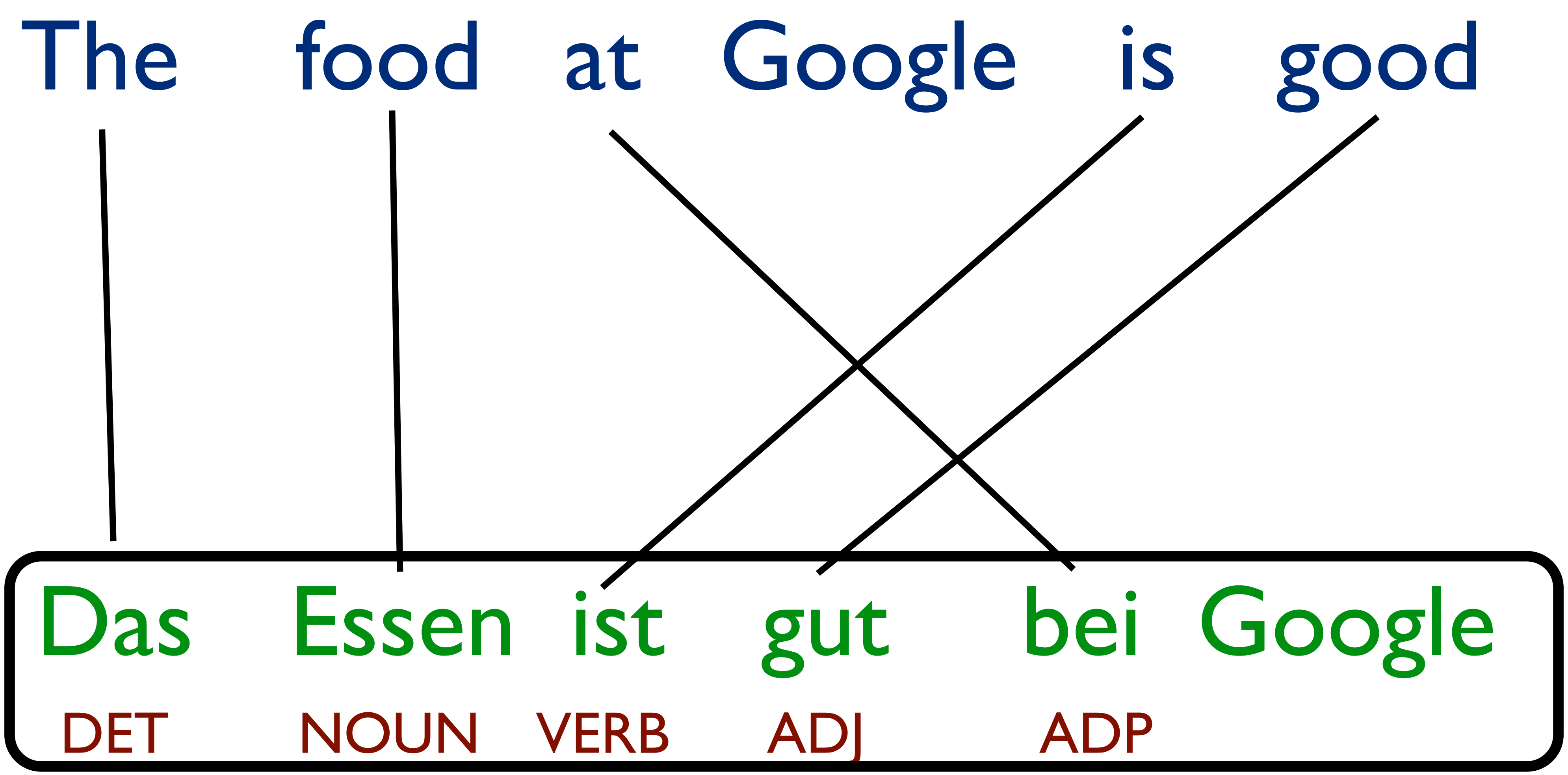
Cross-Lingual Projection



Cross-Lingual Projection

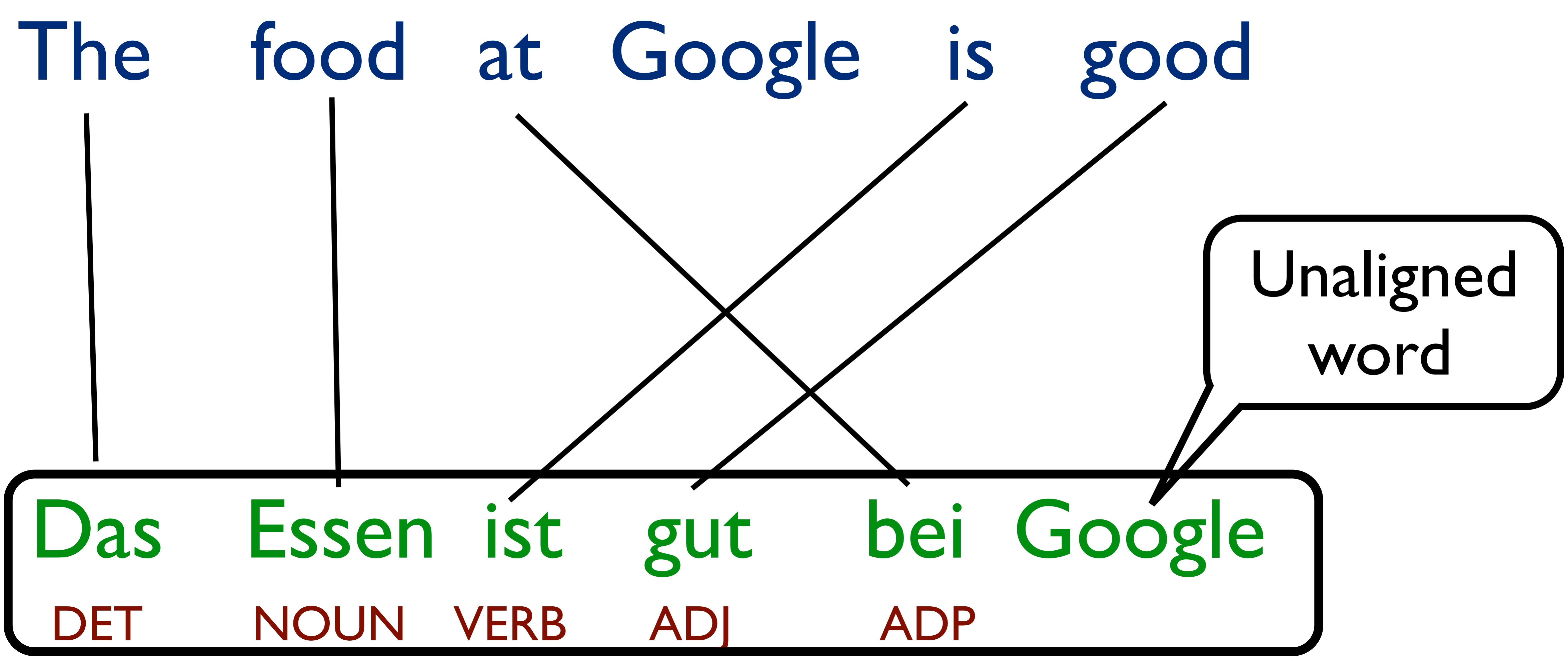


Cross-Lingual Projection



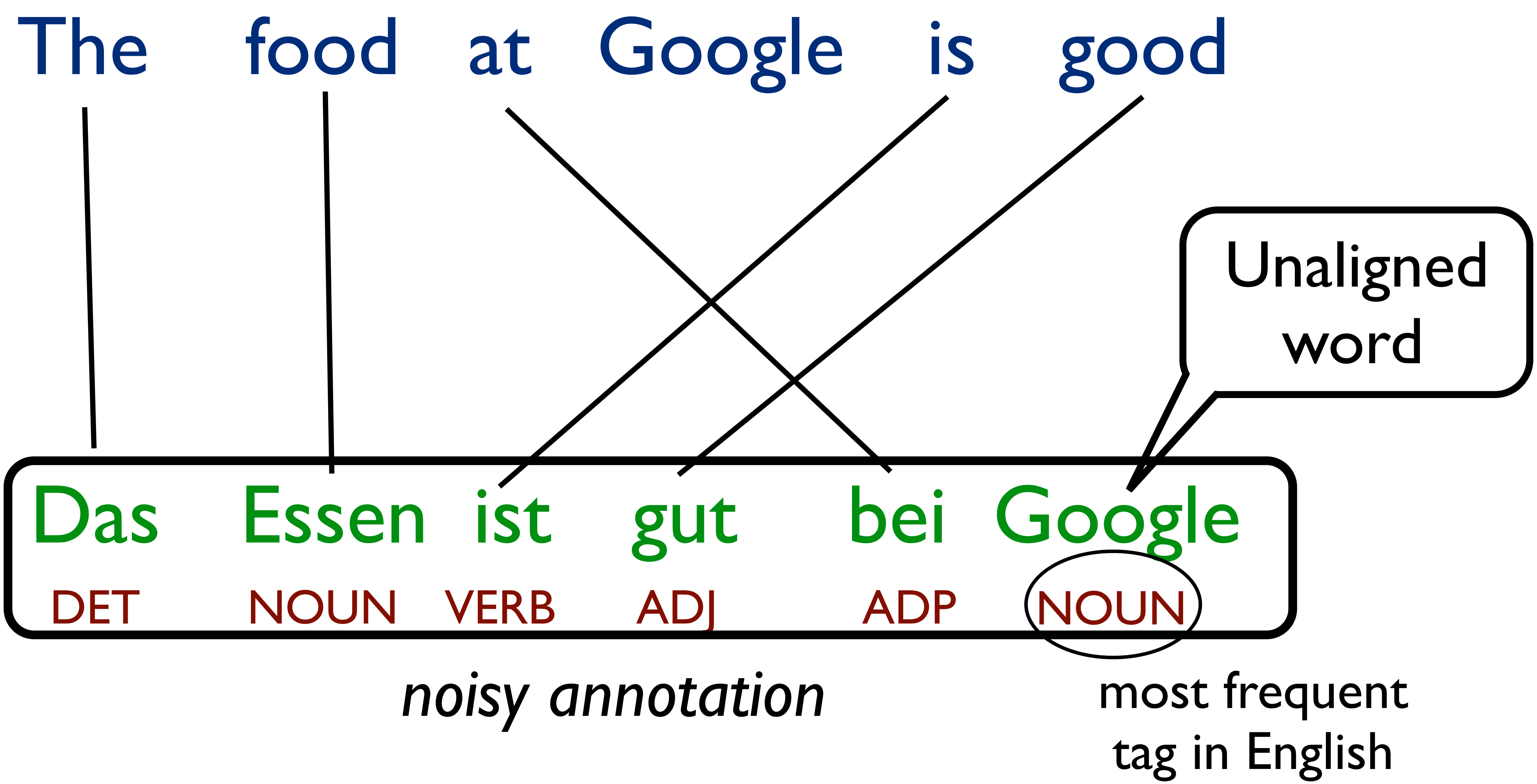
noisy annotation

Cross-Lingual Projection



noisy annotation

Cross-Lingual Projection



Cross-Lingual Projection

Das	Essen	ist	gut	bei	Google
DET	NOUN	VERB	ADJ	ADP	NOUN

+

more projected tagged sentences

supervised training

direct projection **tagger**

Cross-Lingual Projection

Das	Essen	ist	gut	bei	Google
DET	NOUN	VERB	ADJ	ADP	NOUN

\+

more projected tagged sentences

supervised training

direct projection **tagger**

Model 1 : Direct Projection

Model 1 : Direct Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	**73.6**	**77.0**	**83.2**	**79.3**	**79.7**	**82.6**	**80.1**	**74.7**	**78.8**

Yarowsky and Ngai (2001)

Cross-Lingual Projection

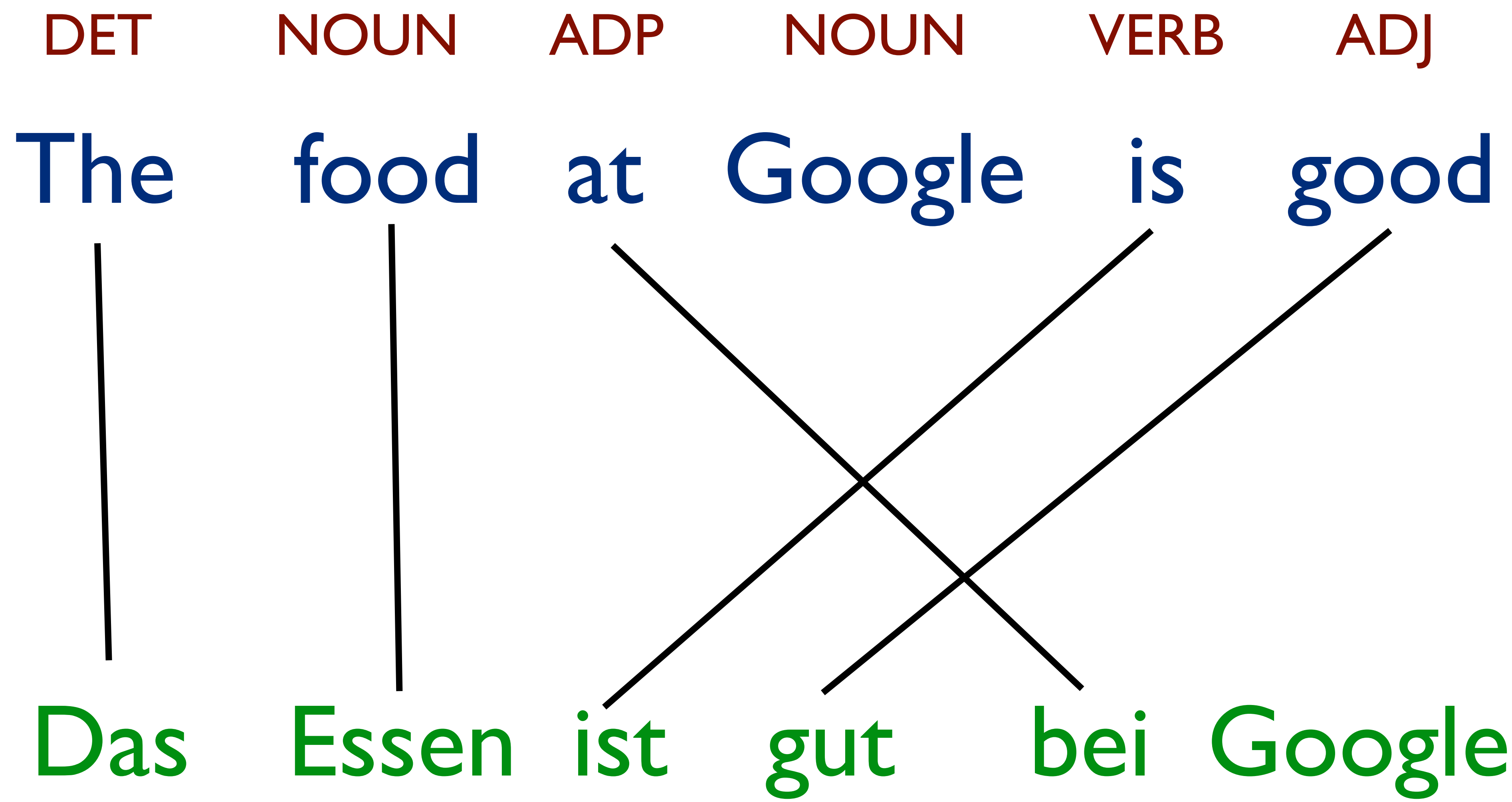
Model 1 : Direct Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	**73.6**	**77.0**	**83.2**	**79.3**	**79.7**	**82.6**	**80.1**	**74.7**	**78.8**

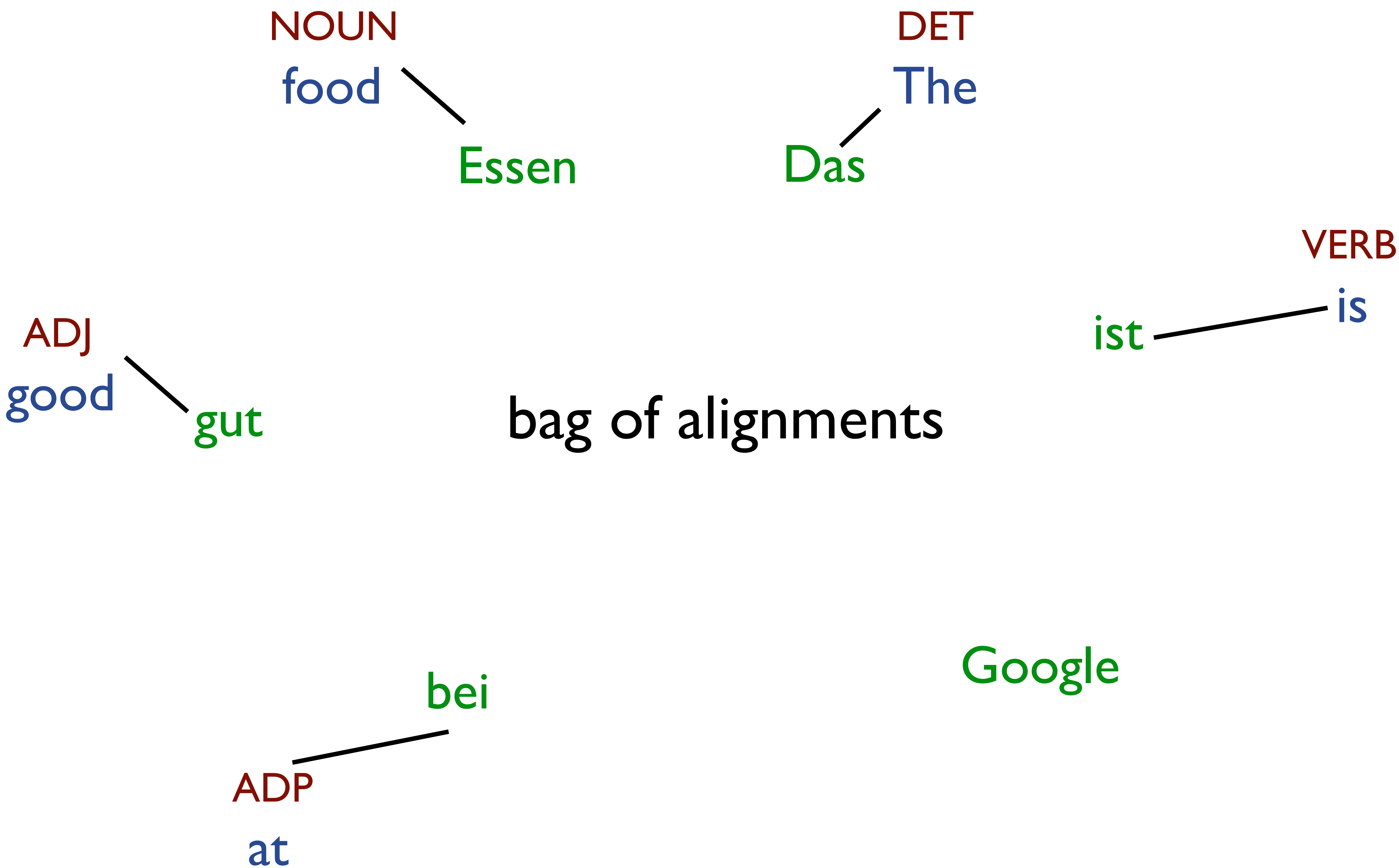
consistent improvements

Yarowsky and Ngai (2001)

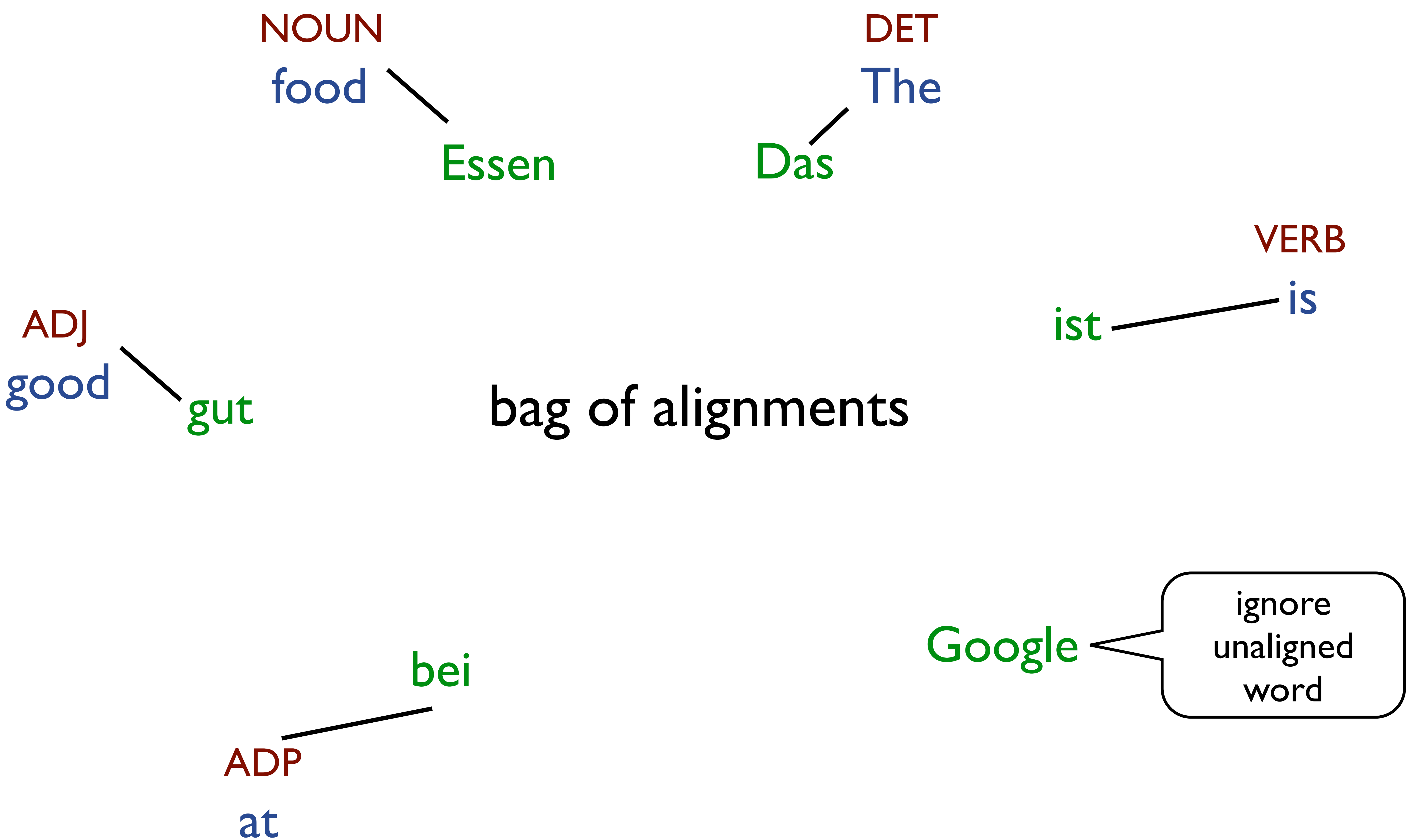
Cross-Lingual Projection



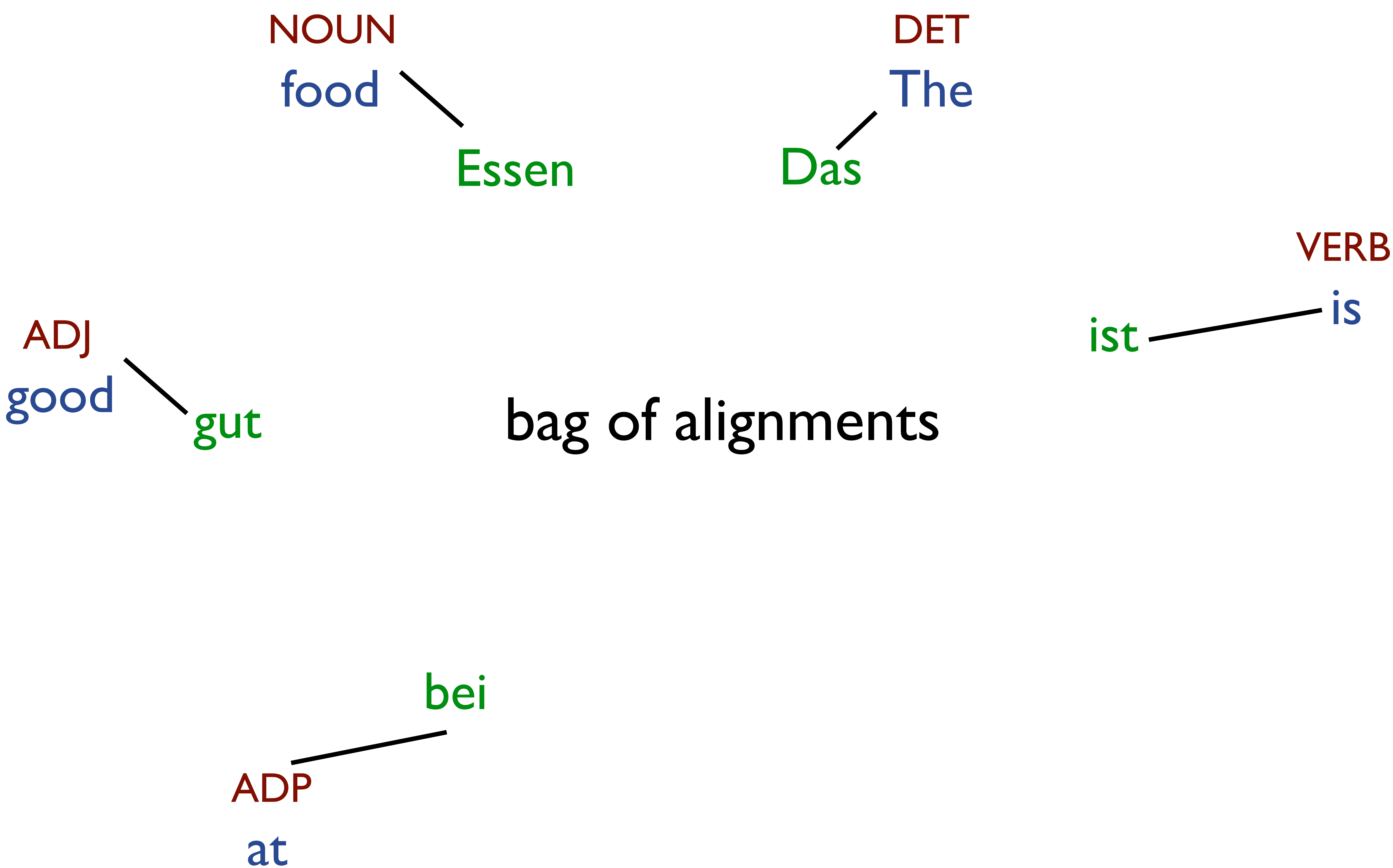
Cross-Lingual Projection



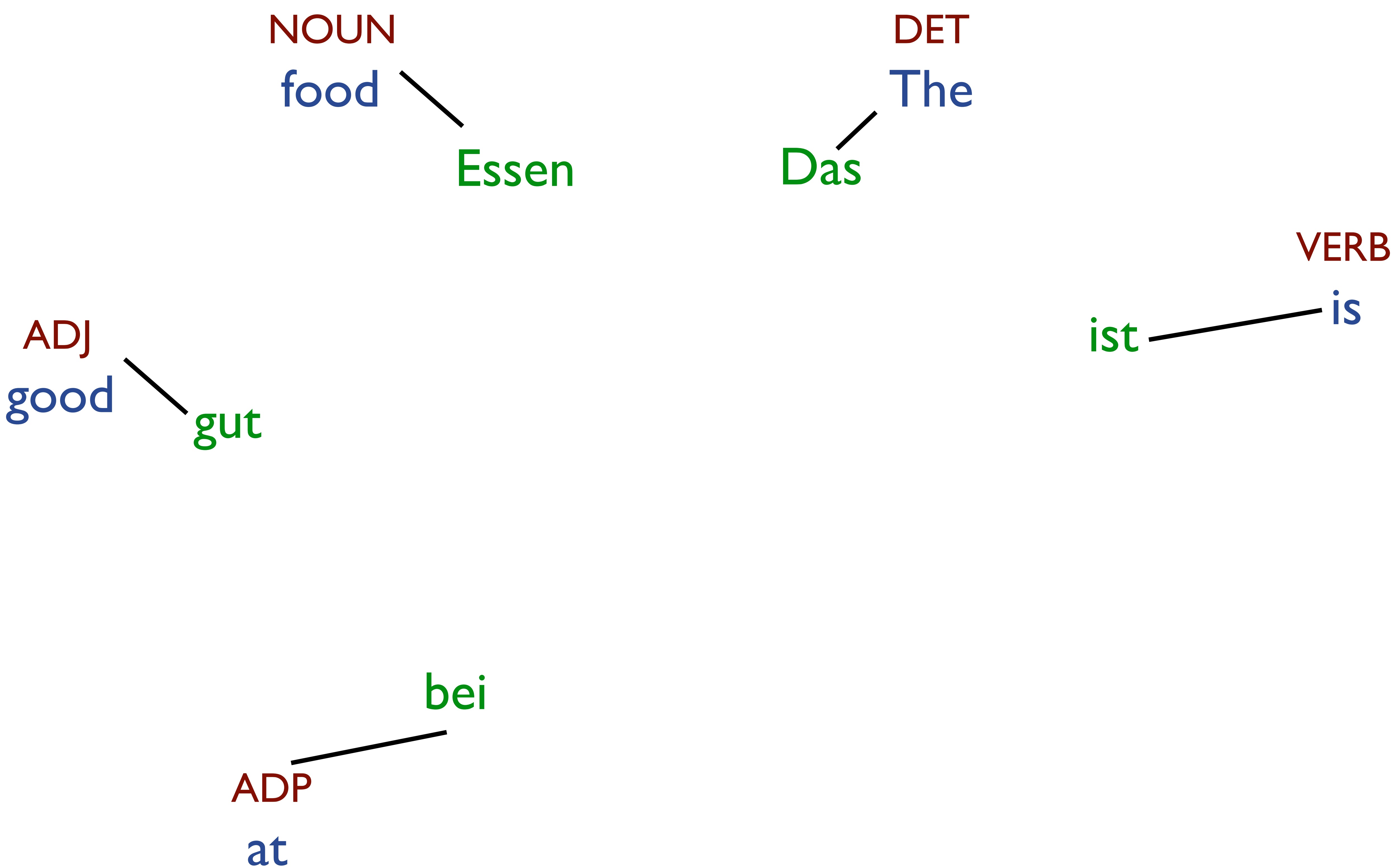
Cross-Lingual Projection



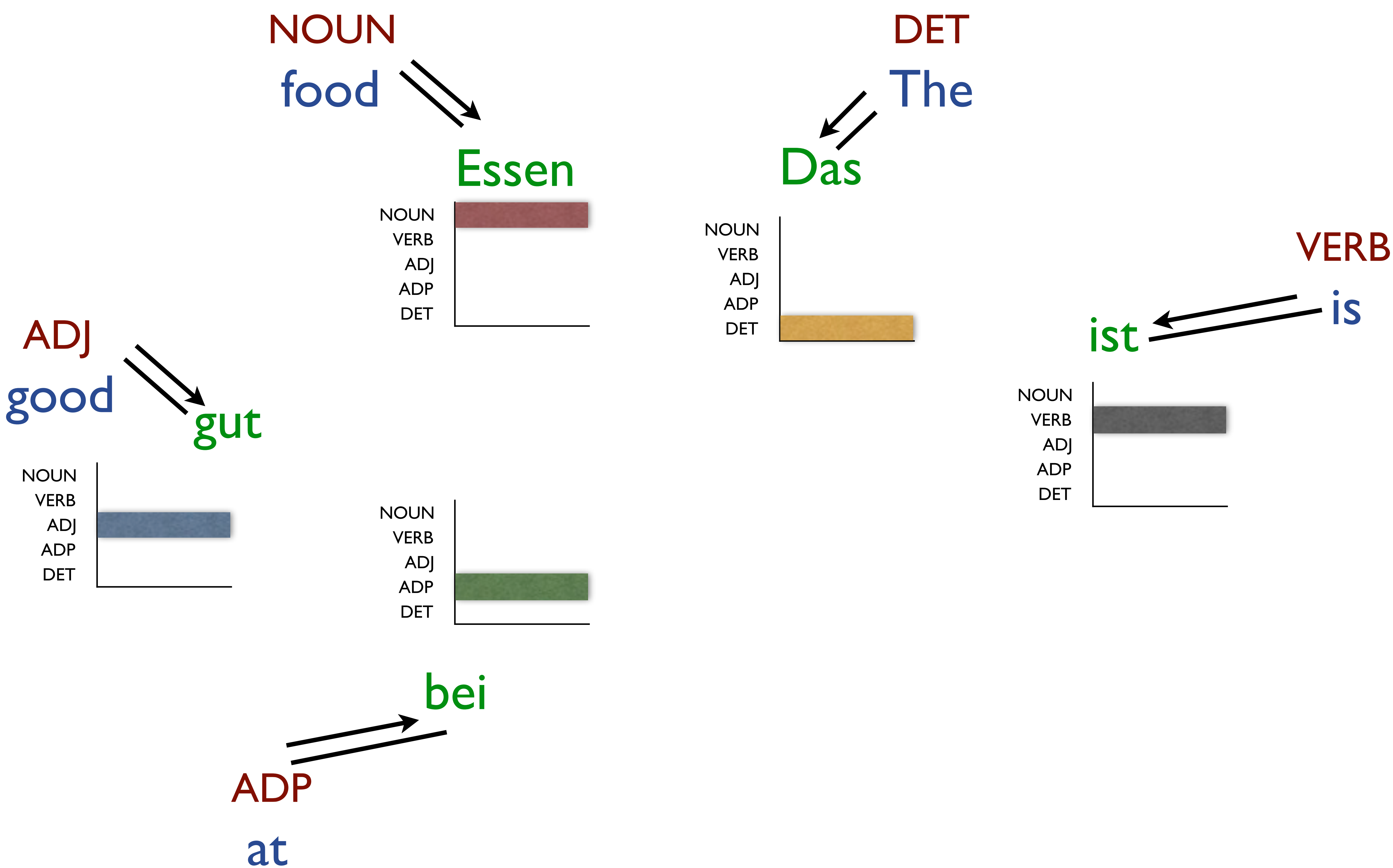
Cross-Lingual Projection



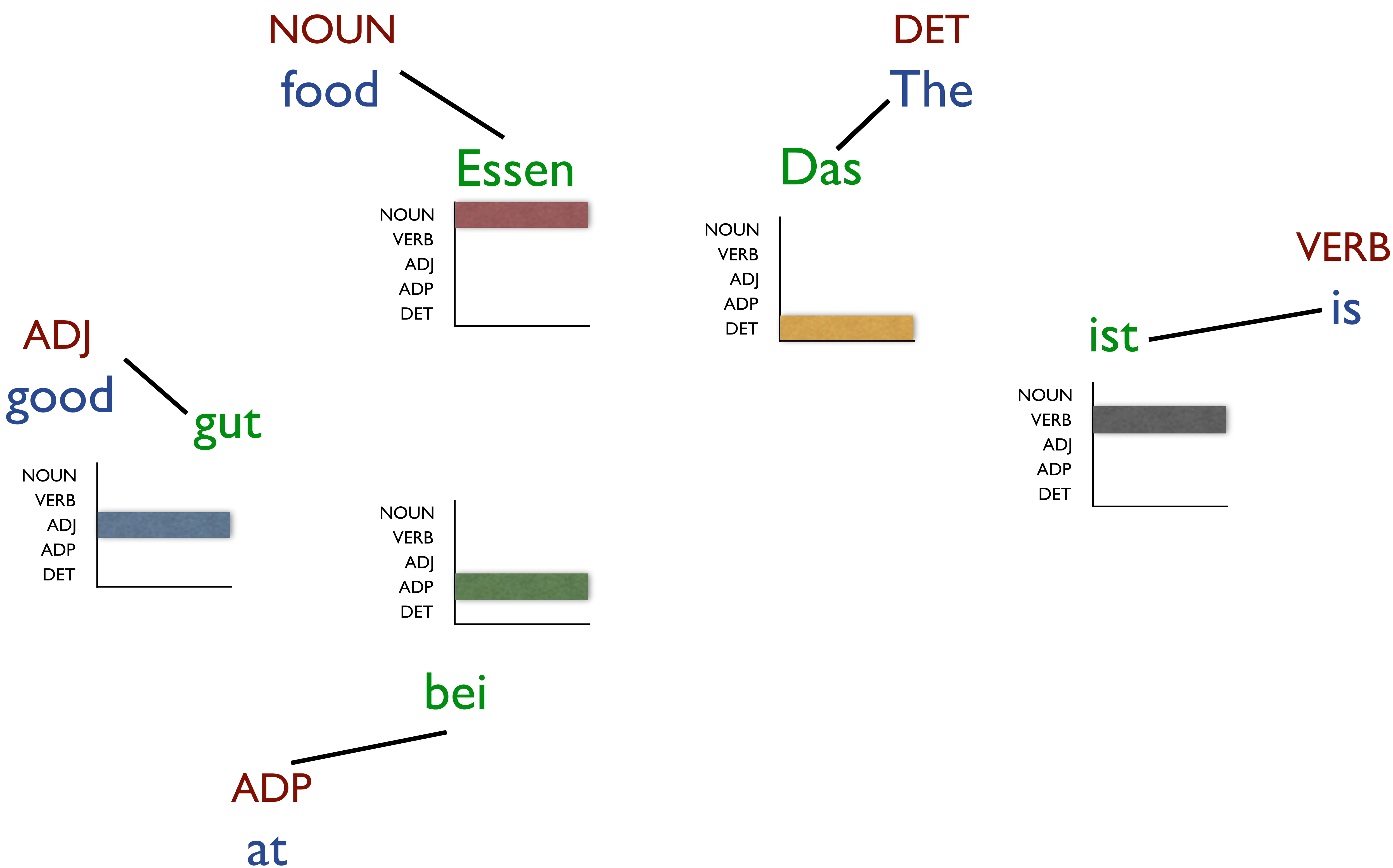
Cross-Lingual Projection



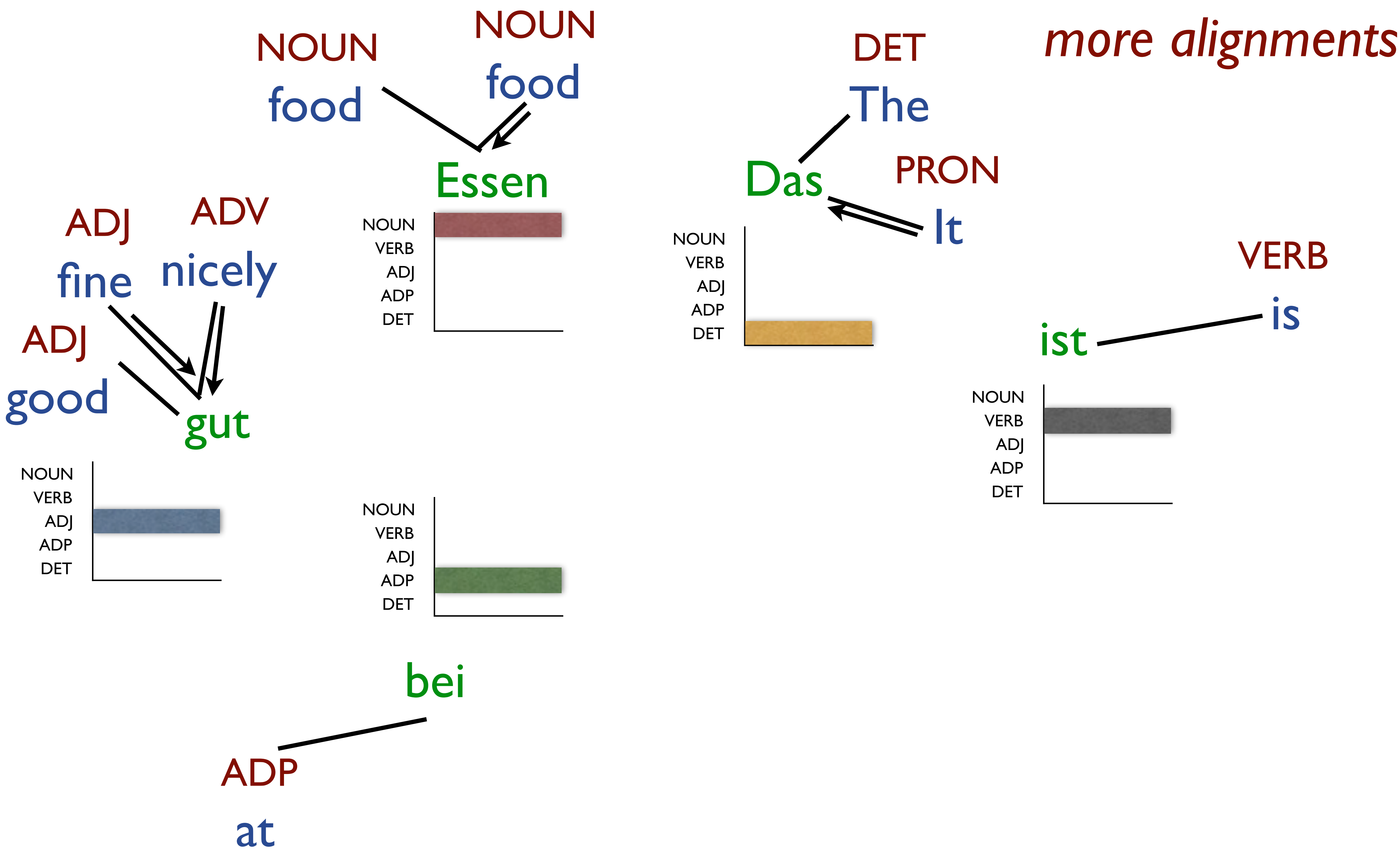
Cross-Lingual Projection



Cross-Lingual Projection

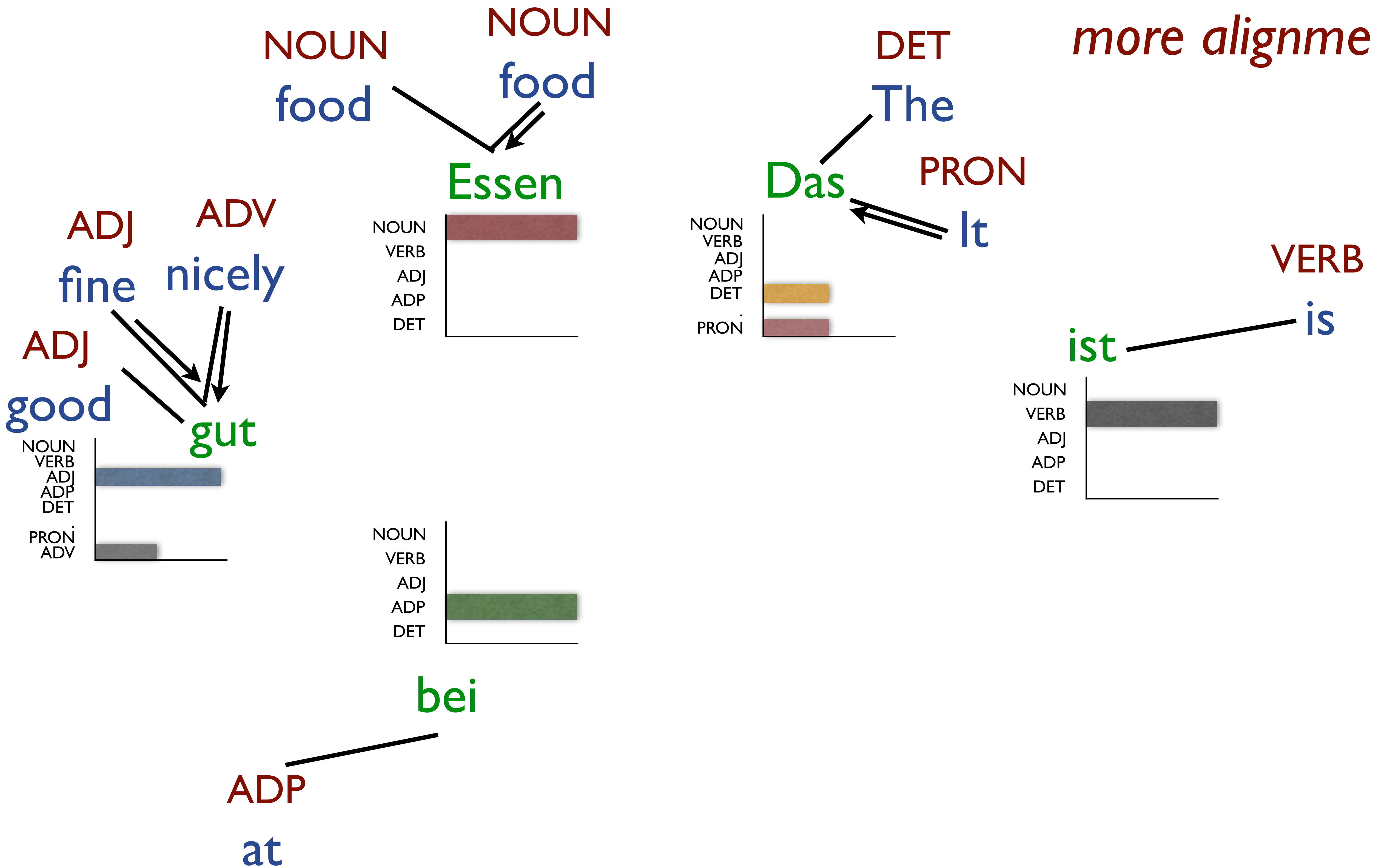


Cross-Lingual Projection



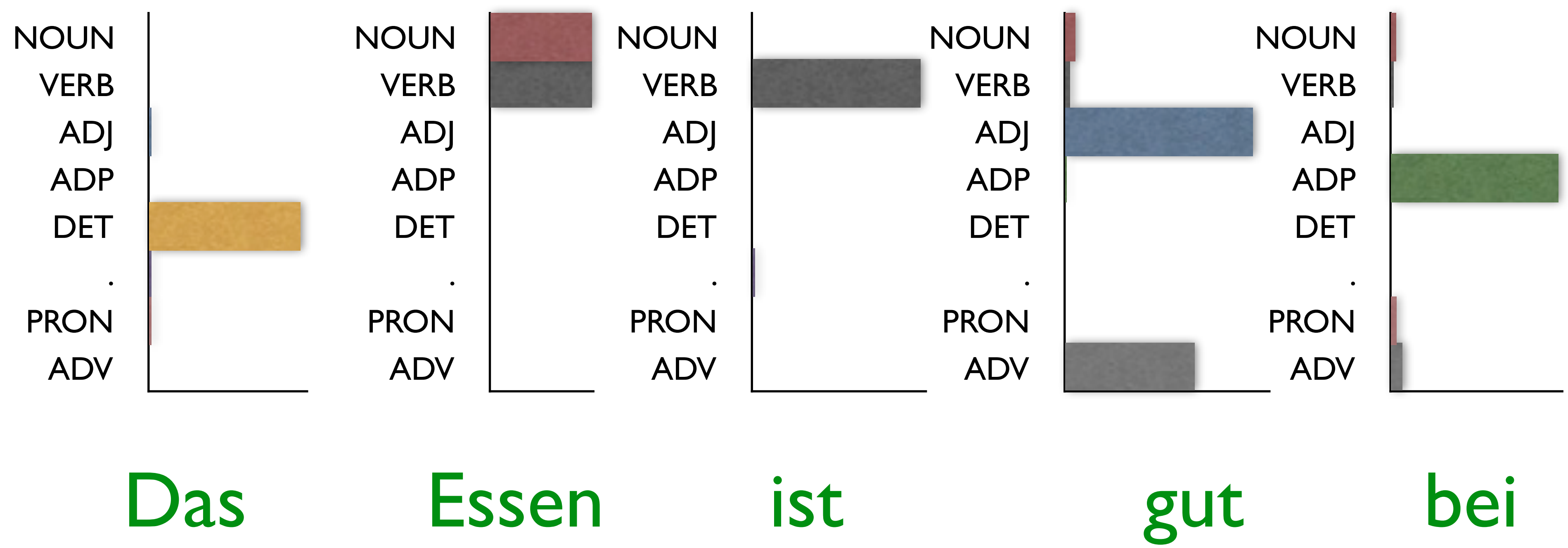
Cross-Lingual Projection

more alignments



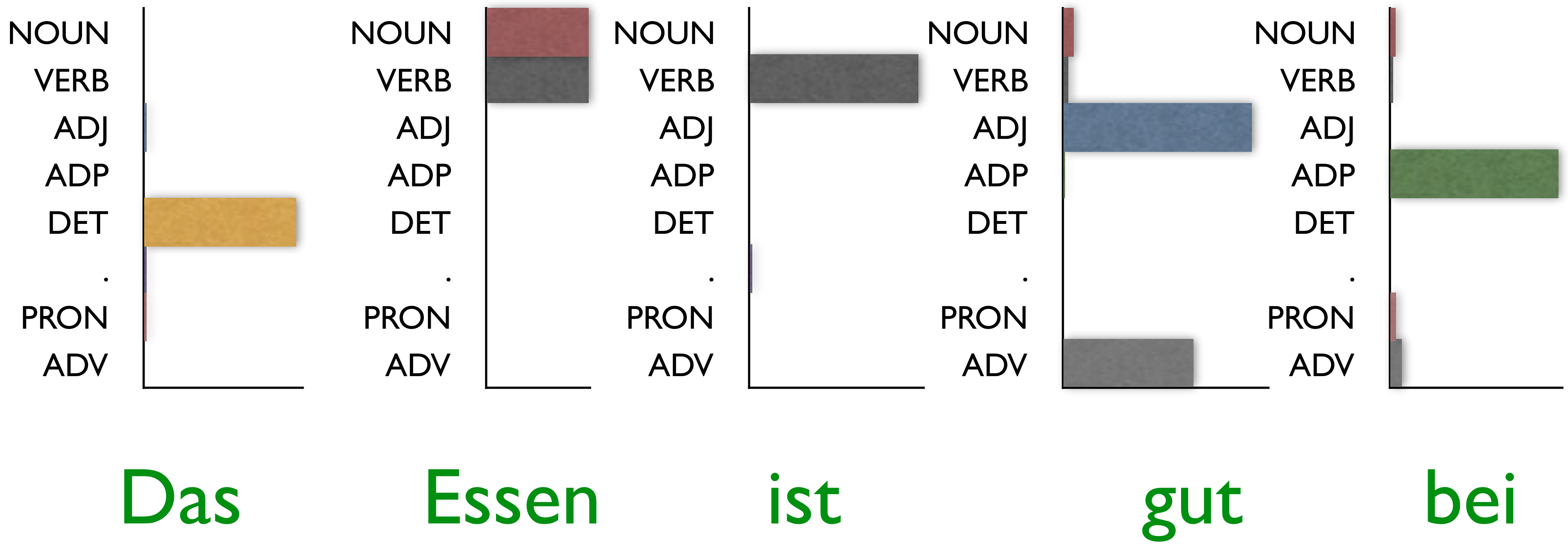
Cross-Lingual Projection

After scanning all the parallel data:



Cross-Lingual Projection

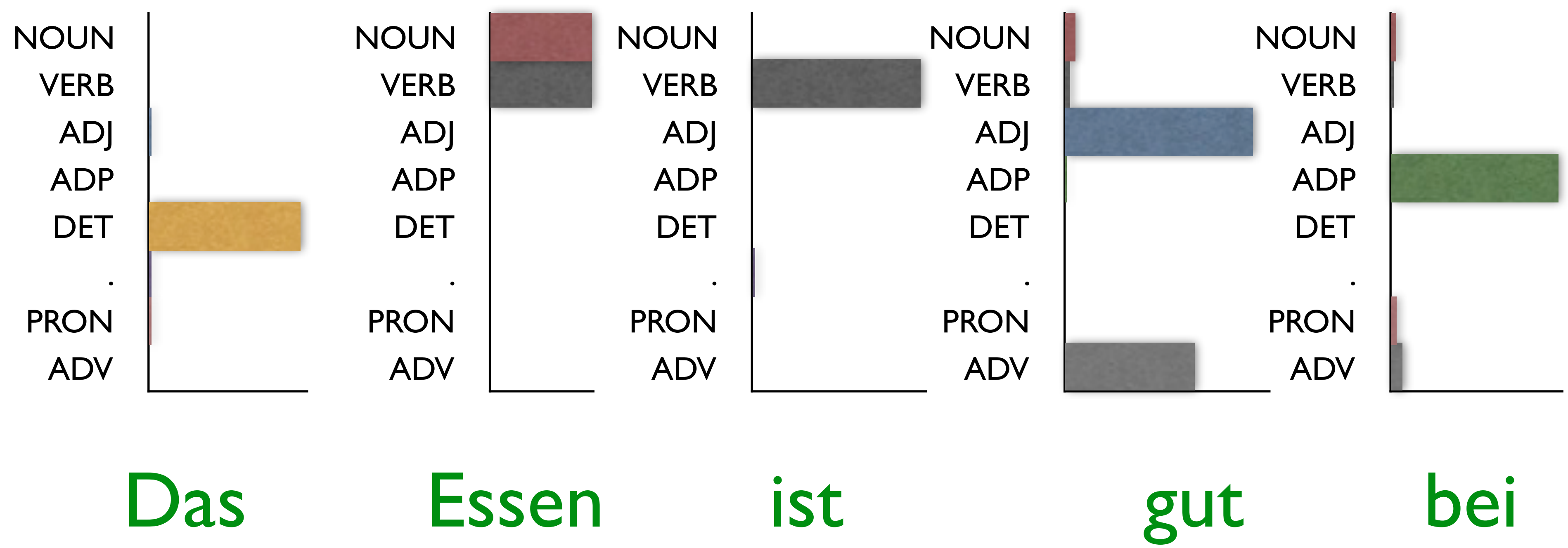
After scanning all the parallel data:



$p(y|x)$ = probability of a tag given a word

Cross-Lingual Projection

After scanning all the parallel data:

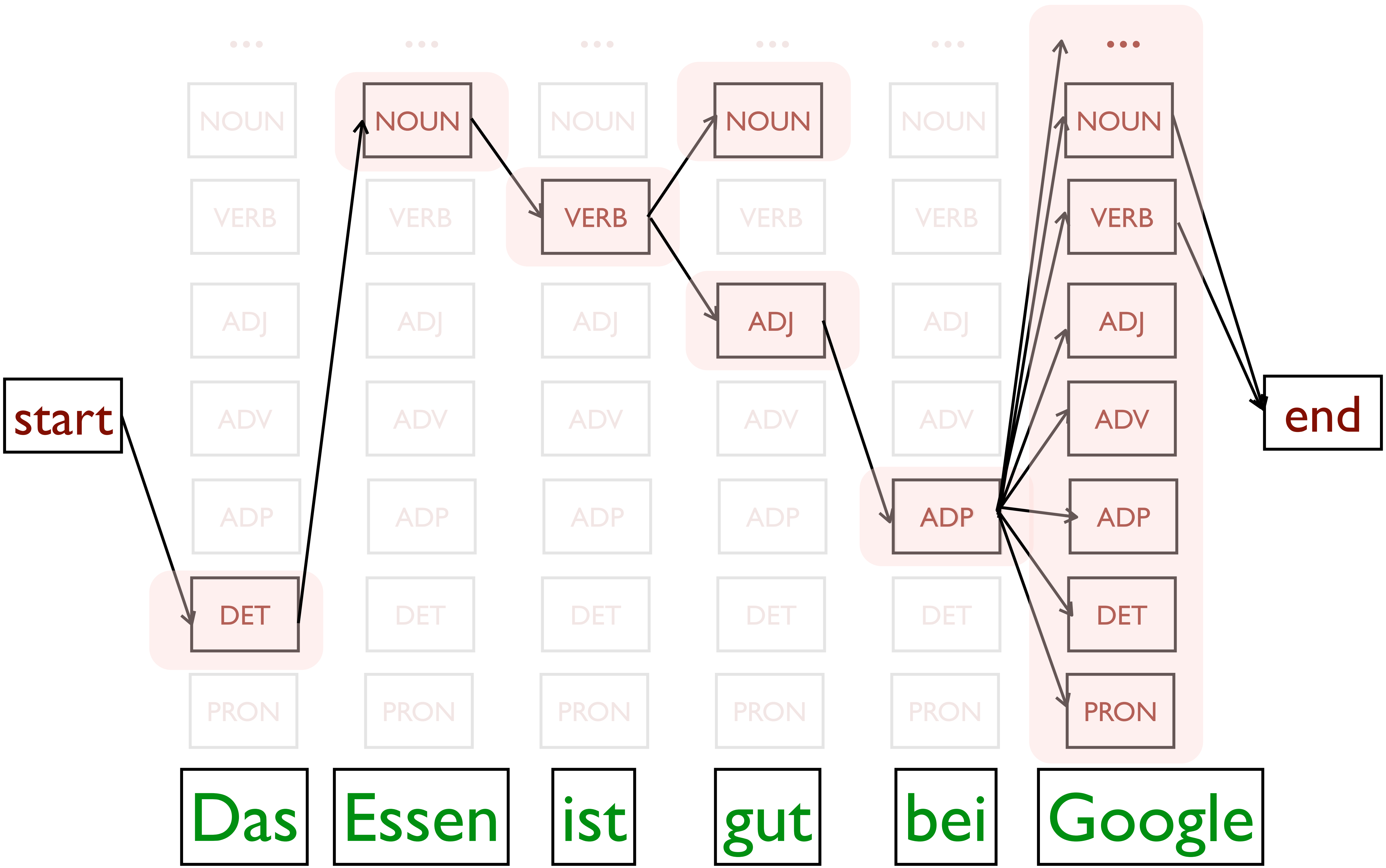


$p(y|x)$ = probability of a tag given a word

$$\boxed{\mathrm{Dict}(x) = \{y : p(y|x) > \tau\}}$$

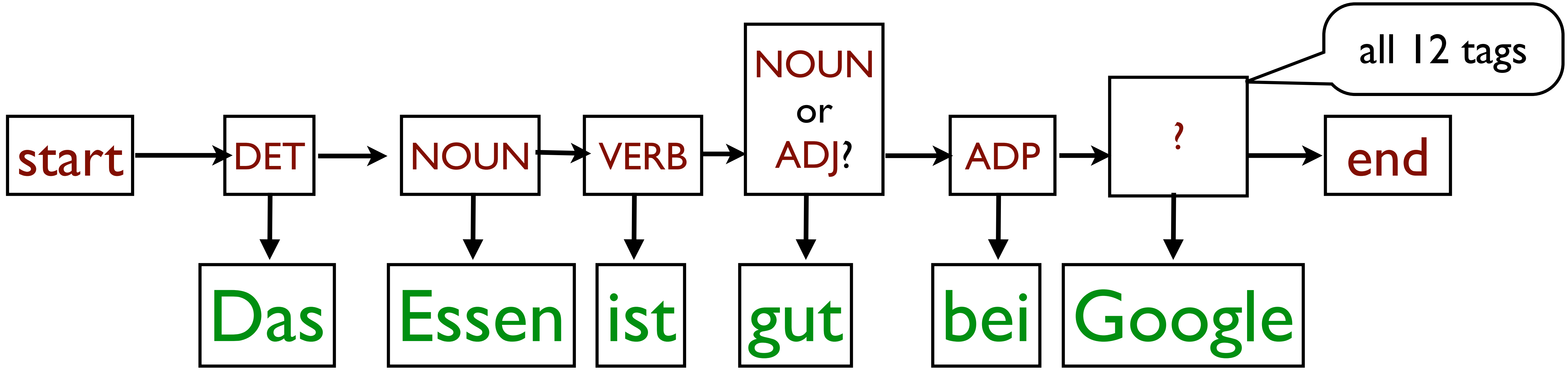
Cross-Lingual Projection

HMM with locally-normalized log-linear models + projected lexicon constraints

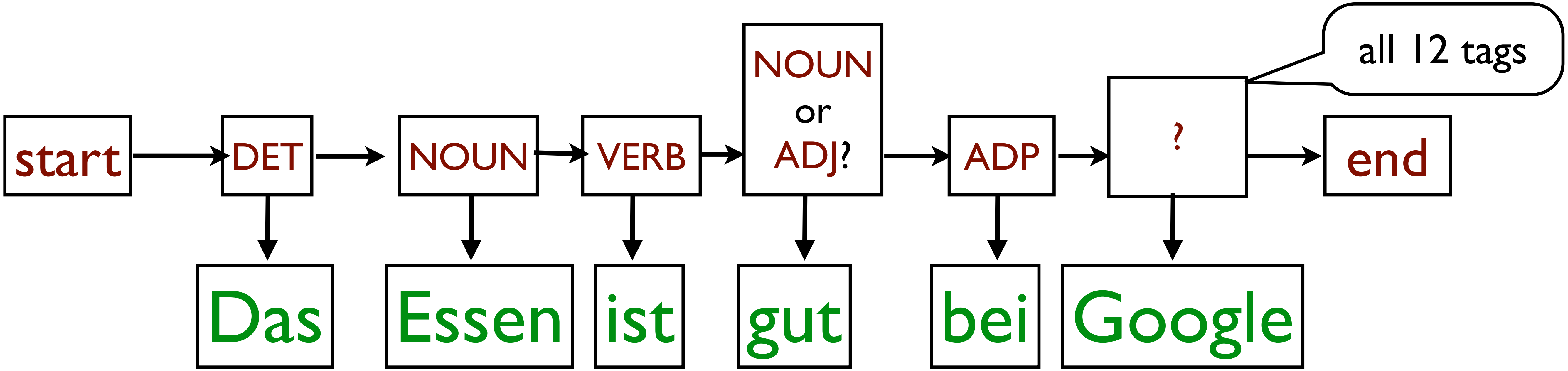


Unsupervised POS Tagging with **Projected Dictionaries**

HMM with locally-normalized log-linear models + projected lexicon constraints



HMM with locally-normalized log-linear models + projected lexicon constraints



Model 2 : Lexicon Projection

HMM with locally-normalized log-linear models + projected lexicon constraints

Model 2 : Lexicon Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	73.6	77.0	**83.2**	**79.3**	79.7	82.6	80.1	74.7	78.8
Projected Dictionary	**79.0**	**78.8**	82.4	76.3	**84.8**	**87.0**	**82.8**	**79.4**	**81.3**

HMM with locally-normalized log-linear models + projected lexicon constraints

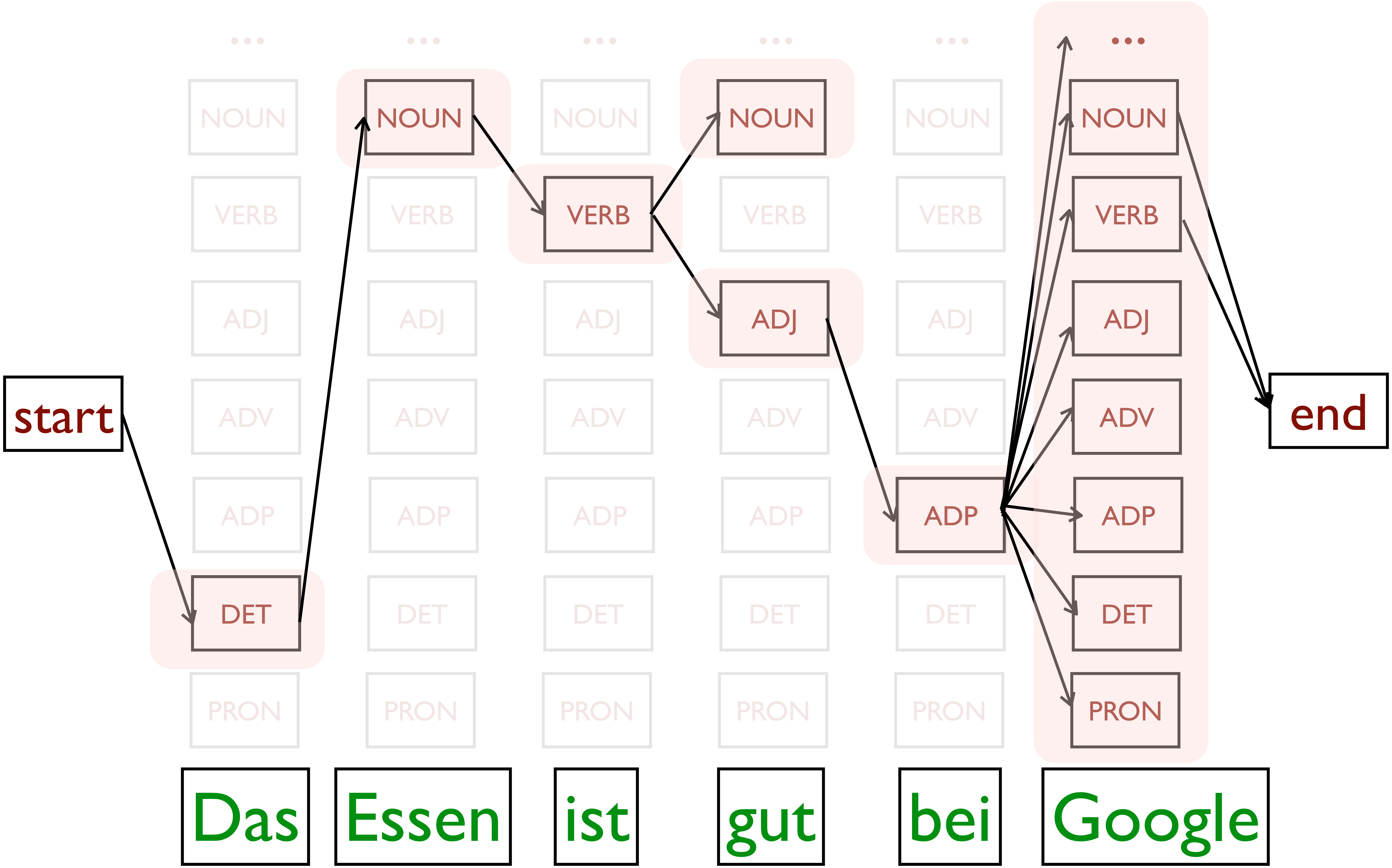
Model 2 : Lexicon Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	73.6	77.0	**83.2**	**79.3**	79.7	82.6	80.1	74.7	78.8
Projected Dictionary	**79.0**	**78.8**	82.4	76.3	**84.8**	**87.0**	**82.8**	**79.4**	**81.3**

Improvement over direct projection for most languages

Unsupervised POS Tagging with **Projected Dictionaries**

HMM with locally-normalized log-linear models + projected lexicon constraints

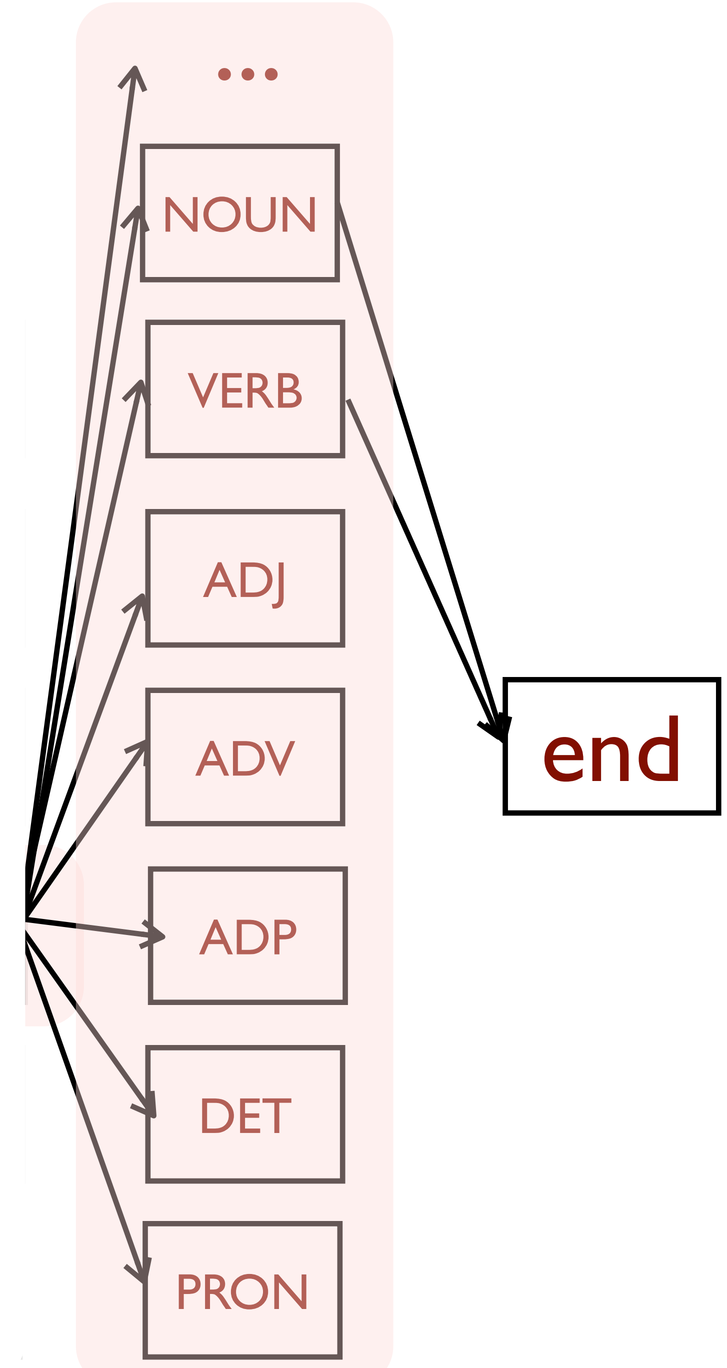


HMM with locally-normalized log-linear models + projected lexicon constraints

Words with no alignments are not covered by the projected dictionaries

Can we automatically improve coverage?

Solution:
graph-based semi-supervised learning



Das Essen ist gut bei Google

Graph-Based Semi-Supervised Learning

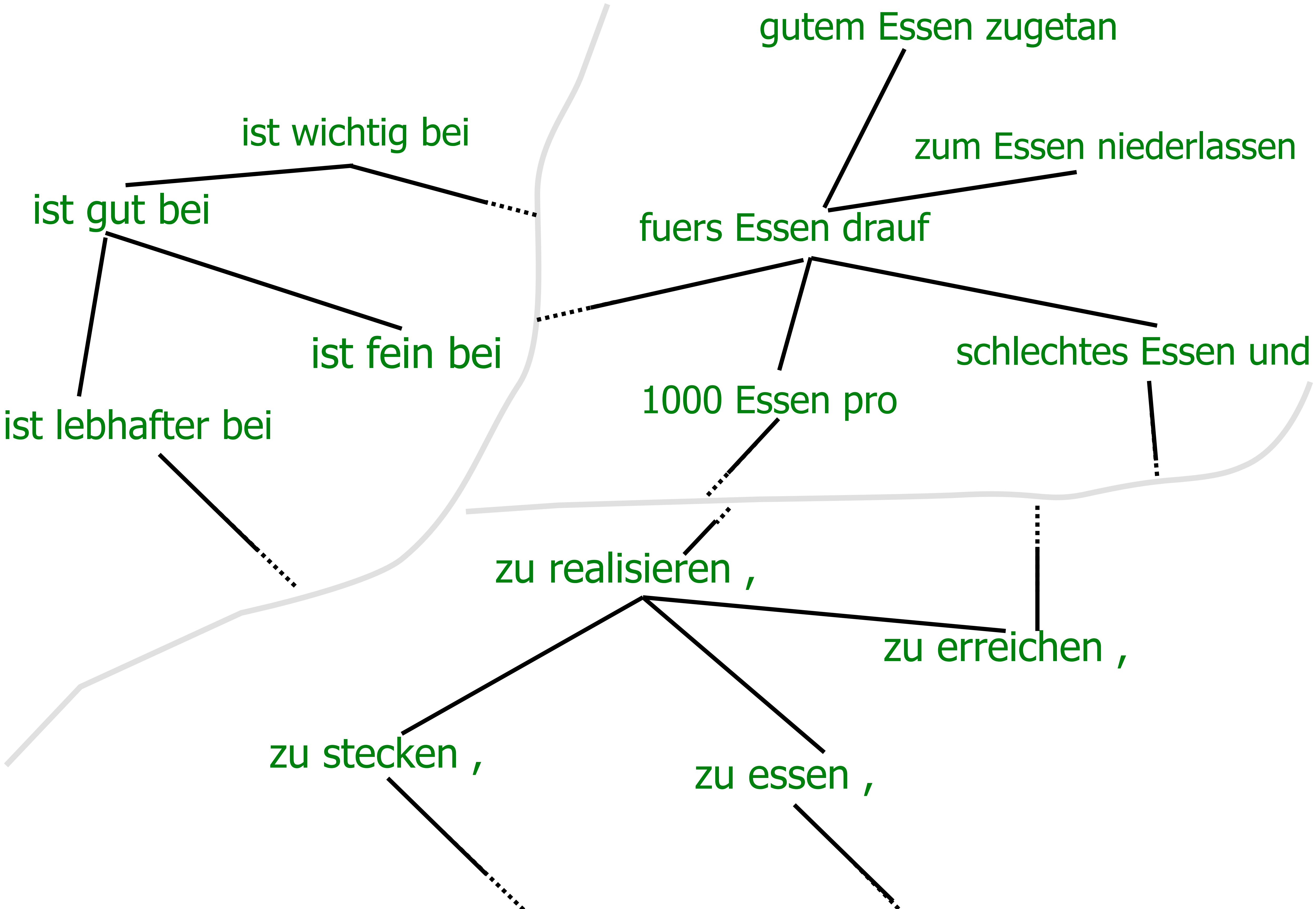
Graph-Based Semi-Supervised Learning

- For a **target** language,
  - Build a graph over several million trigrams types as vertices
  - Compute similarity of vertices using distributional statistics

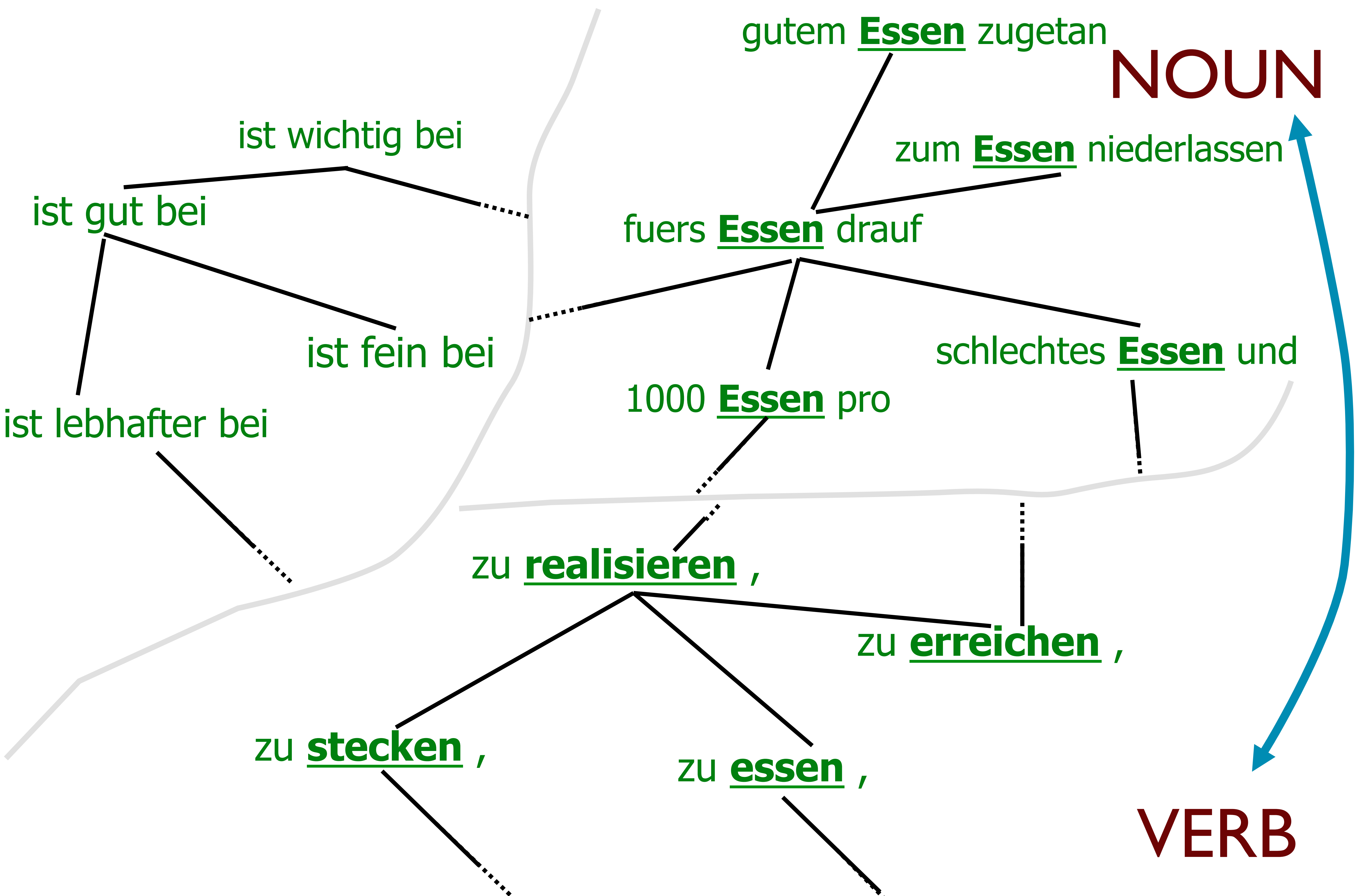
Graph-Based Semi-Supervised Learning

- For a **target** language,
  - Build a graph over several million trigrams types as vertices
  - Compute similarity of vertices using distributional statistics
- Label distribution at each vertex
  - POS tag distribution of the middle word in the trigram

Example Graph in German



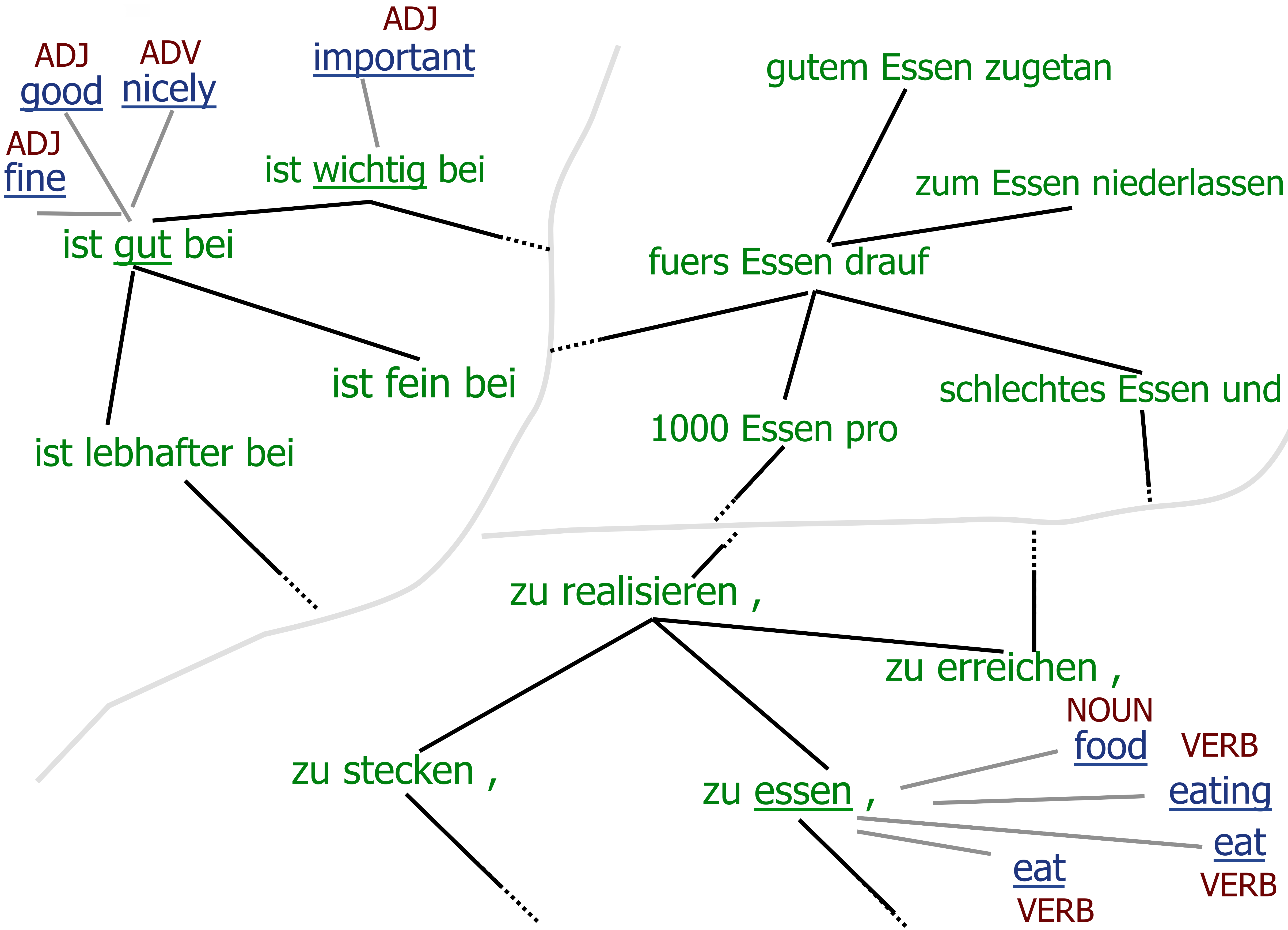
Example Graph in German



Graph-Based Semi-Supervised Learning

- For a **target** language,
  - Build a graph over several million trigrams types as vertices
  - Compute similarity of vertices using distributional statistics
- Label distribution at each vertex
  - POS tag distribution of the middle word in the trigram
- Plug in auto-tagged words from a **source** language
- Links between source and target vertices are word alignments

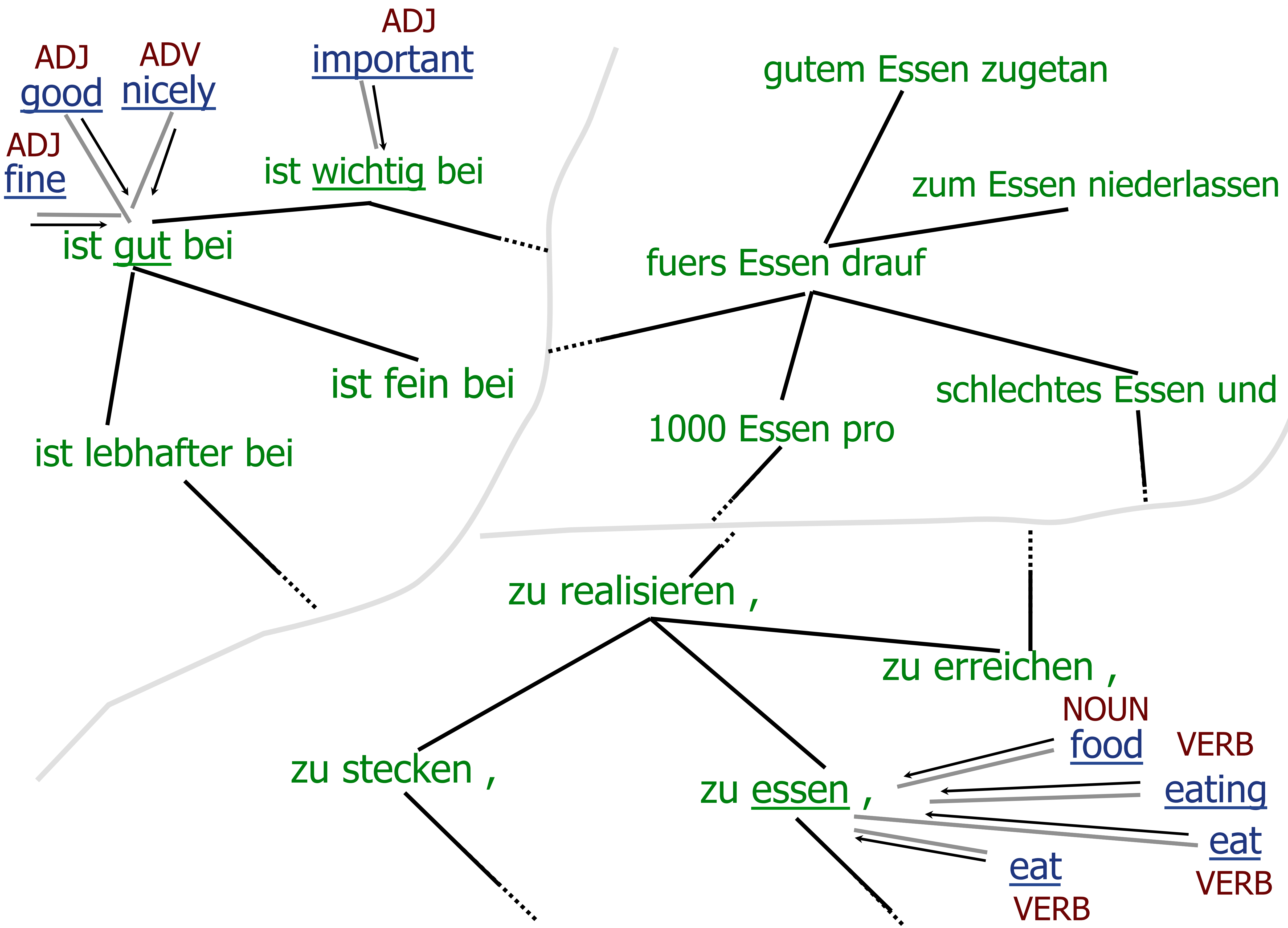
Bilingual Graph



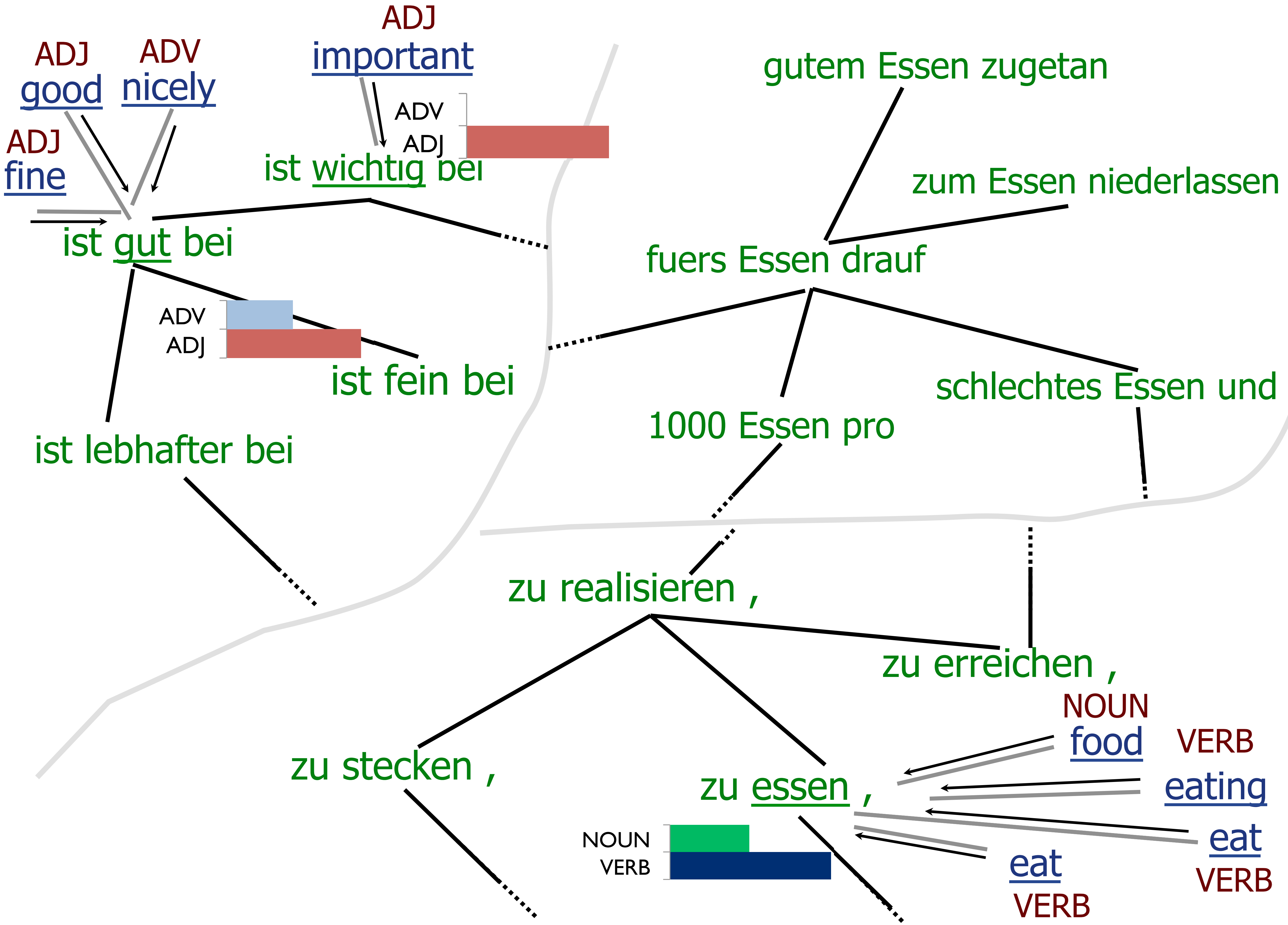
Graph-Based Semi-Supervised Learning

- Plug in auto-tagged words from a **source** language
- Links between source and target vertices are word alignments
- Run first stage of *label propagation*
  - Source language → Target language

Bilingual Graph-Based Label Propagation



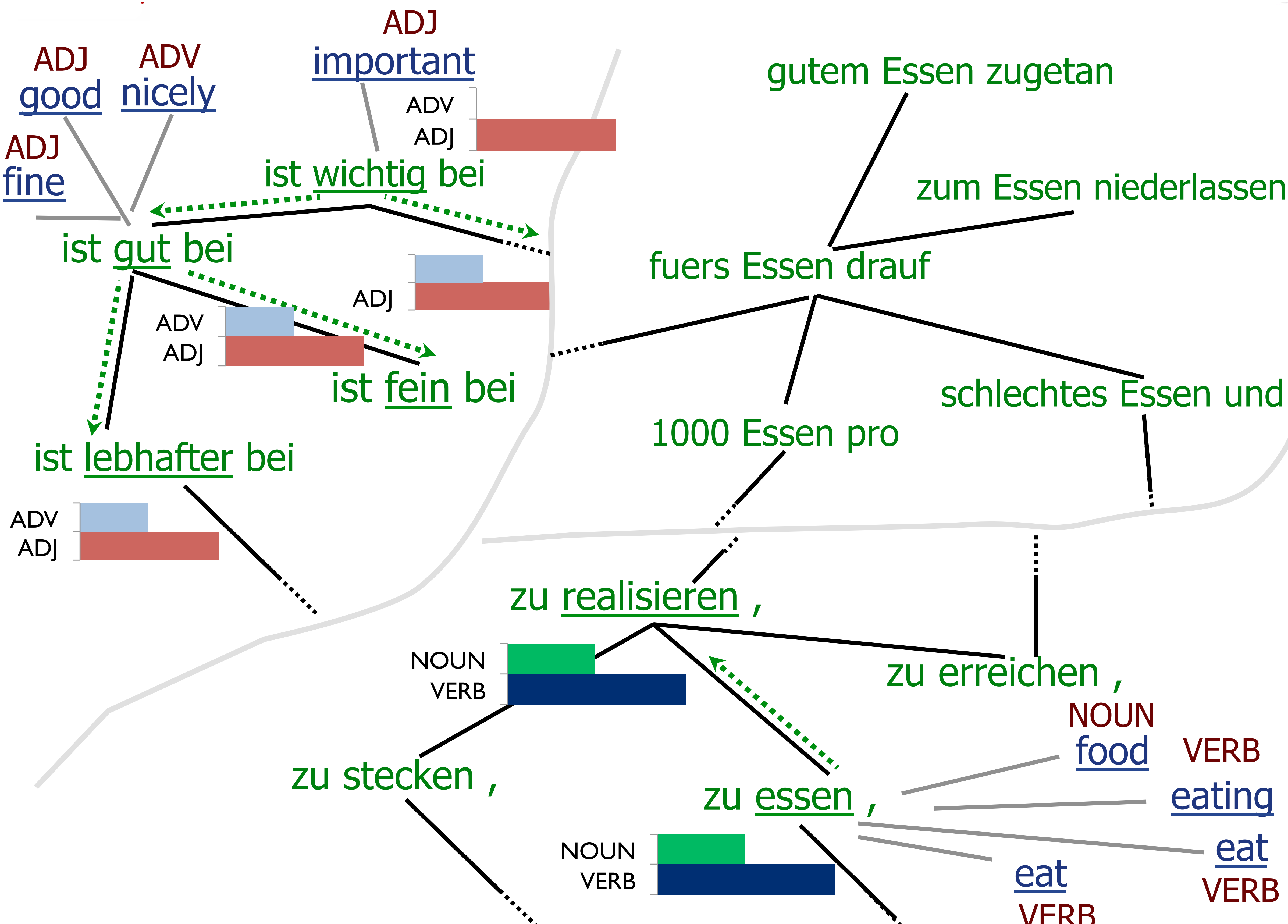
Bilingual Graph-Based Label Propagation



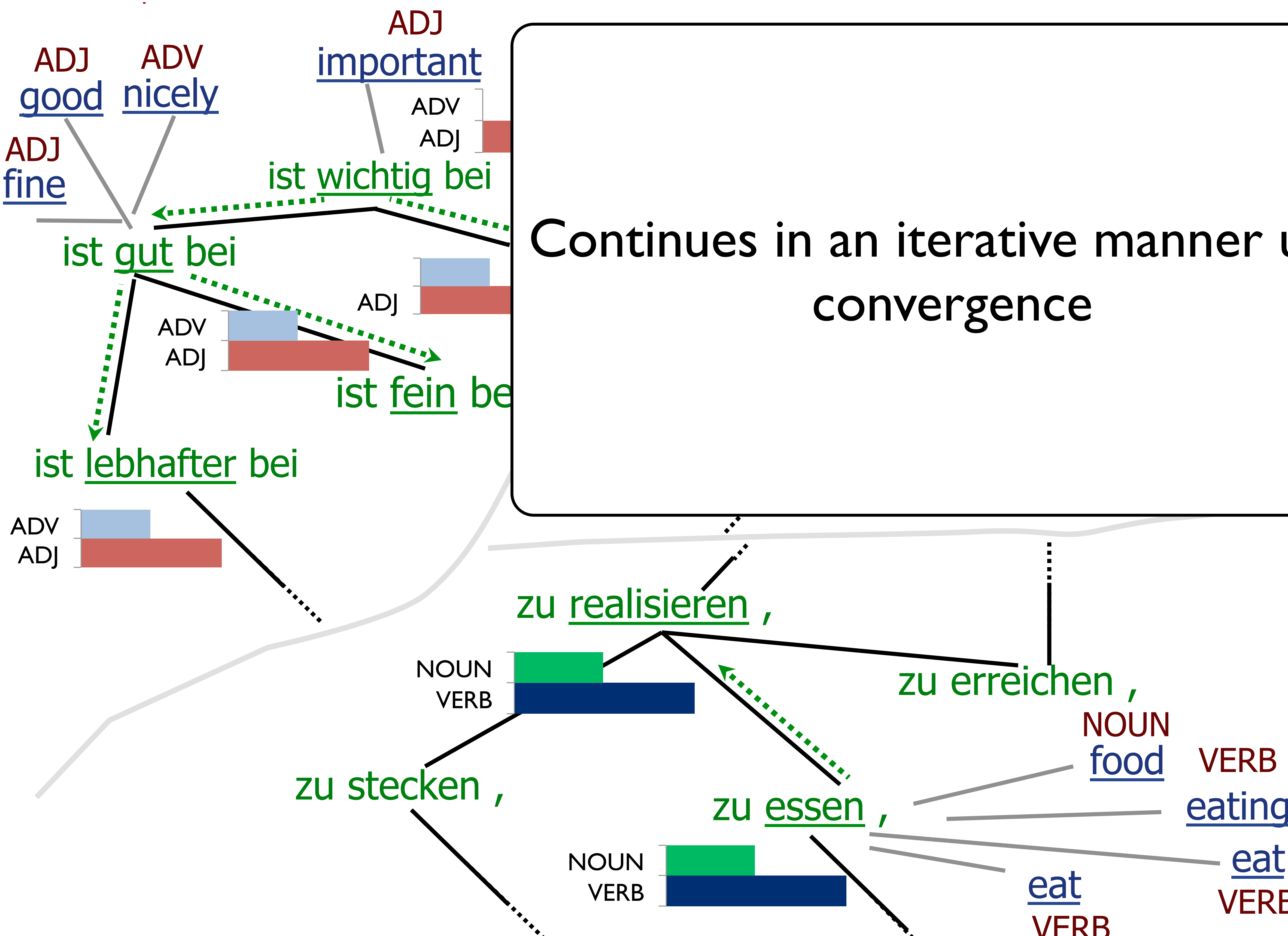
Graph-Based Semi-Supervised Learning

- Plug in auto-tagged words from a **source** language
- Links between source and target vertices are word alignments
- Run first stage of *label propagation*
  - Source language → Target language
- Run second stage of label propagation
  - Within target language vertices
  - Graph objective function with squared penalties

Bilingual Graph Label Propagation



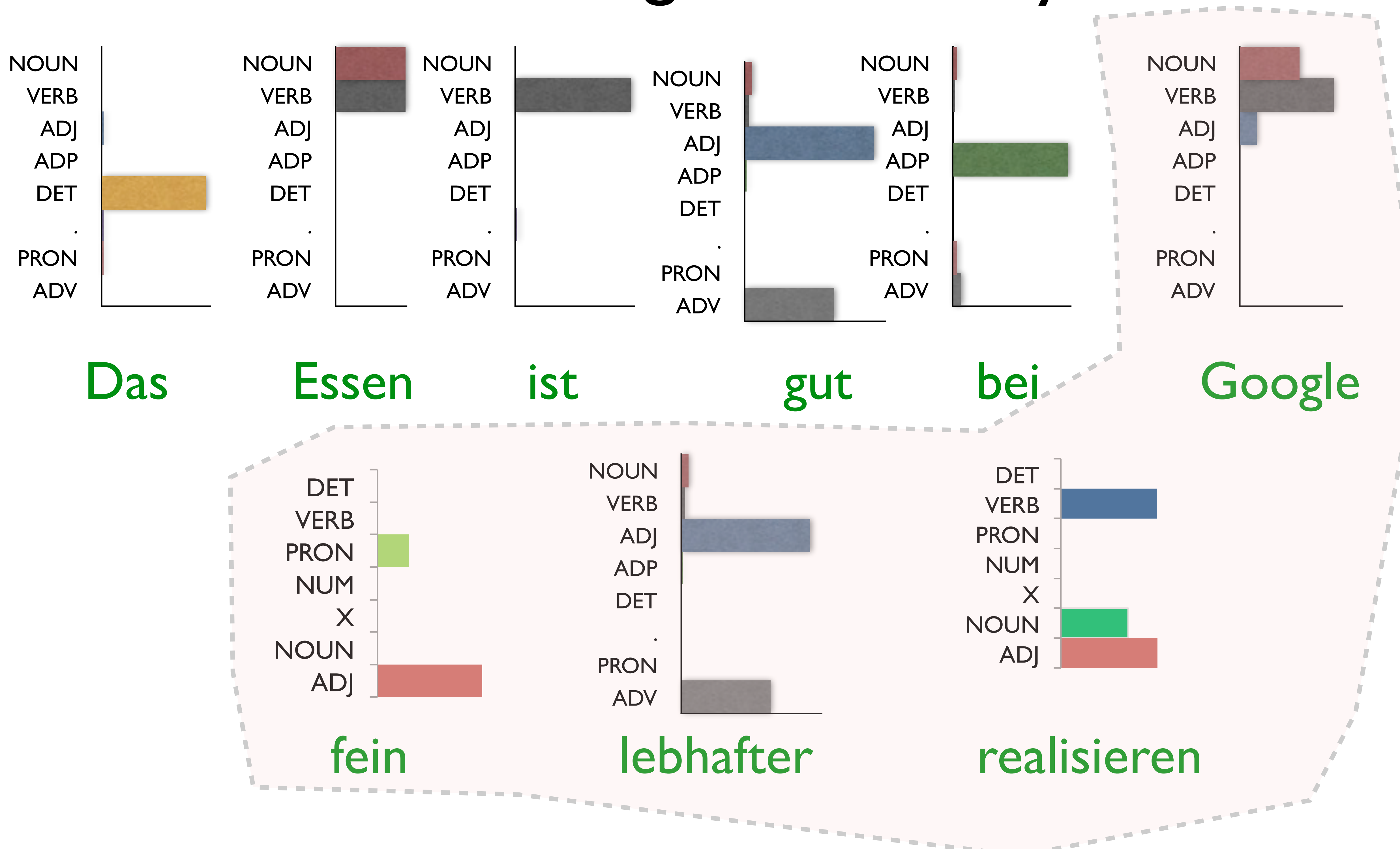
Bilingual Graph Label Propagation



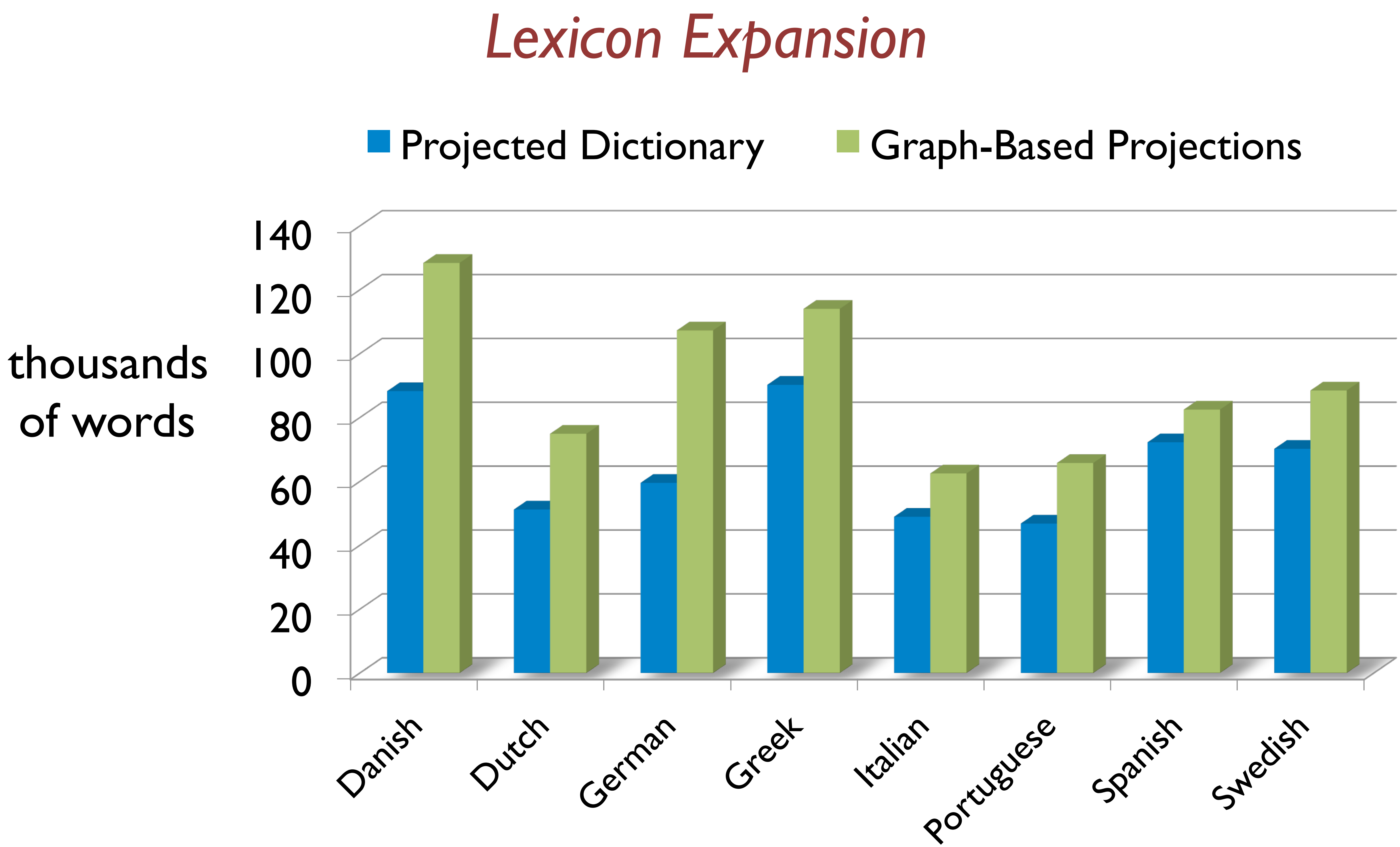
Continues in an iterative manner until convergence

Bilingual Graph-Based Label Propagation

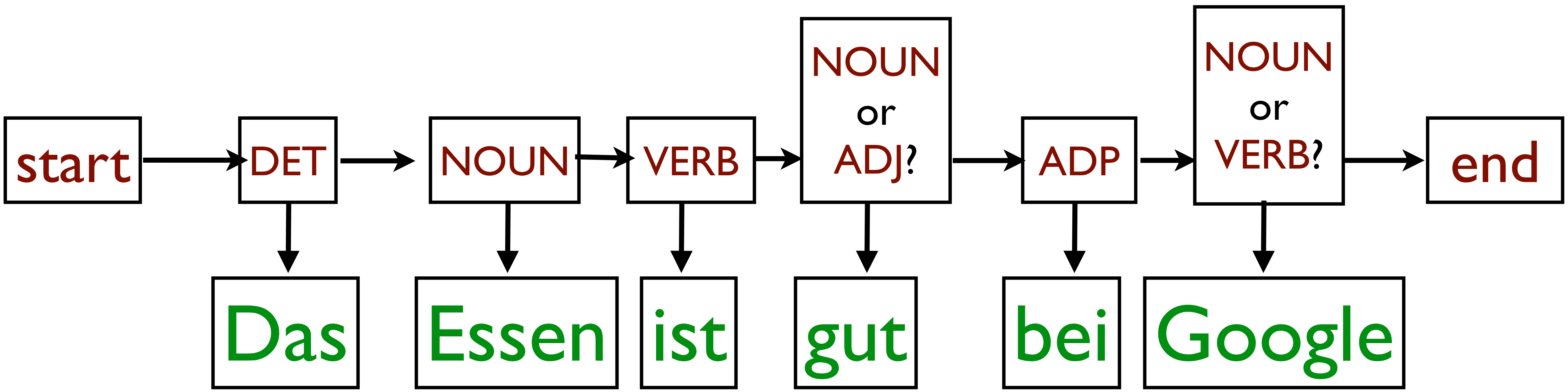
End result : larger dictionary



Bilingual Graph-Based Label Propagation



HMM with locally-normalized log-linear models + graph-based lexicon constraints



Model 3 : Graph-Based Lexicon Projection

HMM with locally-normalized log-linear models + graph-based lexicon constraints

Model 3 : Graph-Based Lexicon Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Projected Dictionary	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
Graph-Based Projections	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**

HMM with locally-normalized log-linear models + graph-based lexicon constraints

Model 3 : Graph-Based Lexicon Projection

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Projected Dictionary	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
Graph-Based Projections	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**
w/ gold dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

Talk Outline

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

Talk Outline

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

So far ...

- Incorporating **type** constraints from a dictionary in a hidden Markov model
  - **type** signifies information from a dictionary or lexicon

So far ...

- Incorporating **type** constraints from a dictionary in a hidden Markov model
  - **type** signifies information from a dictionary or lexicon

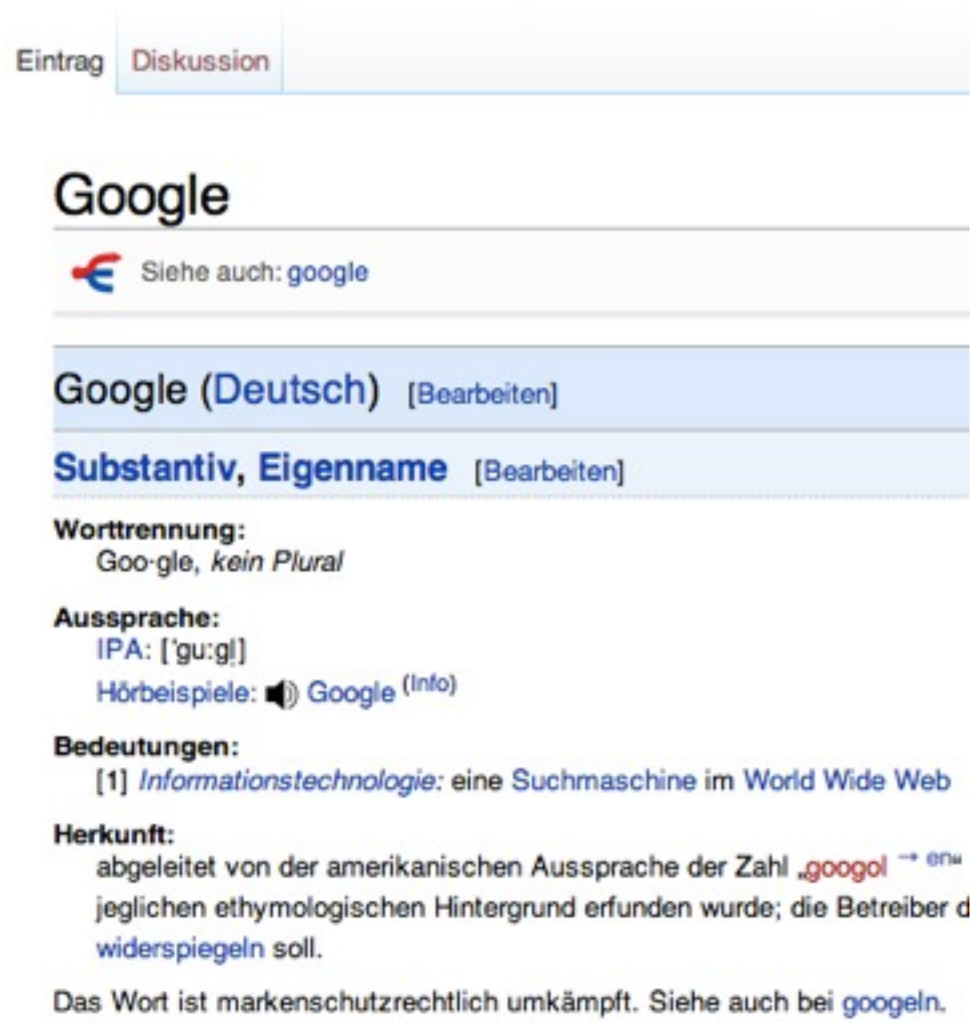
- No **token** level supervision in the previous models
  - Sequence information is important in any tagging problem
  - Can we inject information to our model from parallel data?
  - Can we switch to a more powerful, discriminative model like in supervised learning?

Better Dictionaries

- Recent research, inspired by our work resorted to the use of crowdsourced dictionaries
  - Li, Graça and Taskar (2012)

Better Dictionaries

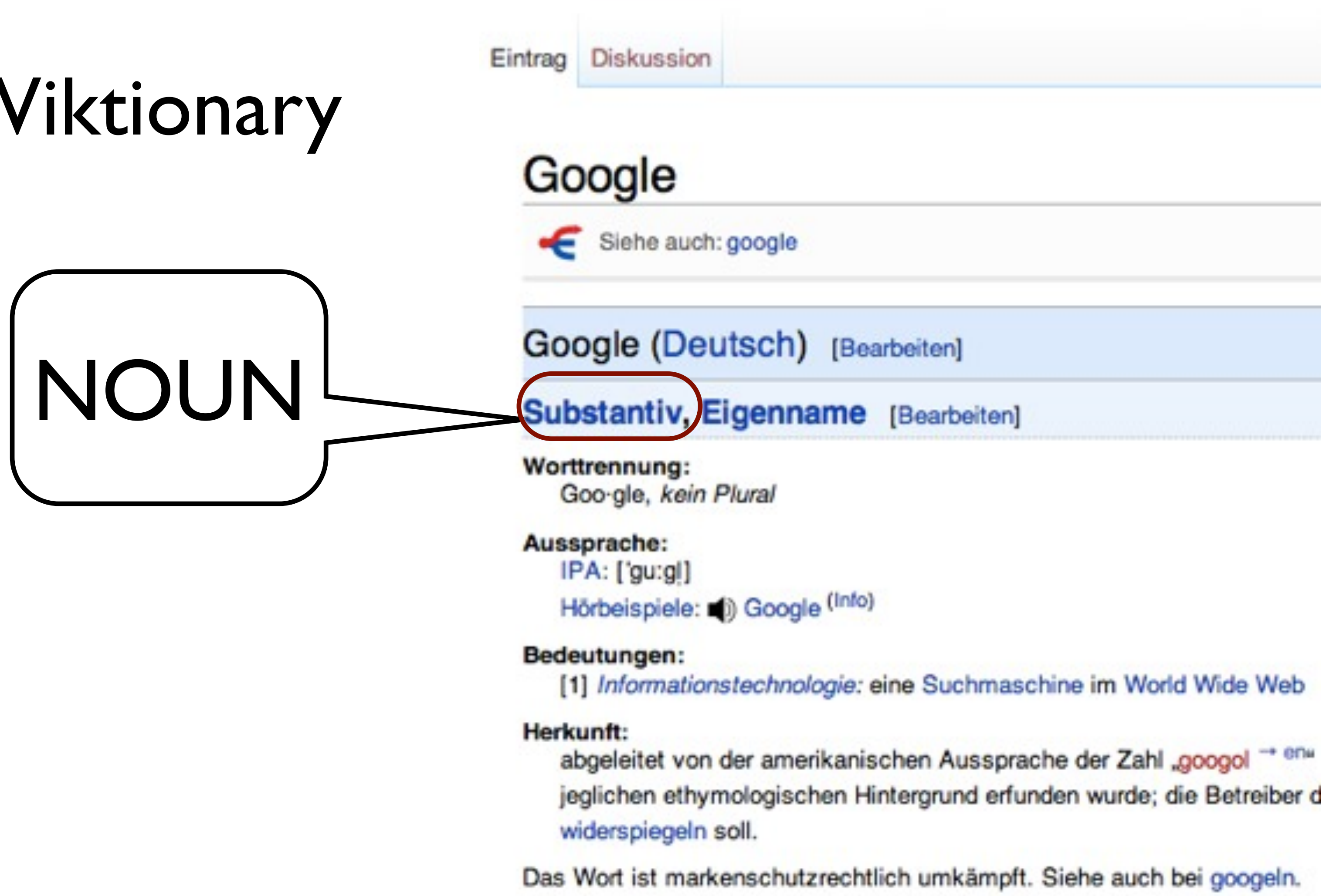
- Recent research, inspired by our work resorted to the use of crowdsourced dictionaries
  - Li, Graça and Taskar (2012)
  - Wiktionary



Better Dictionaries

Google

- Recent research, inspired by our work resorted to the use of crowdsourced dictionaries
  - Li, Graça and Taskar (2012)
  - Wiktionary



Better Dictionaries

- Recent research, inspired by our work resorted to the use of crowdsourced dictionaries
  - Li, Graça and Taskar (2012)

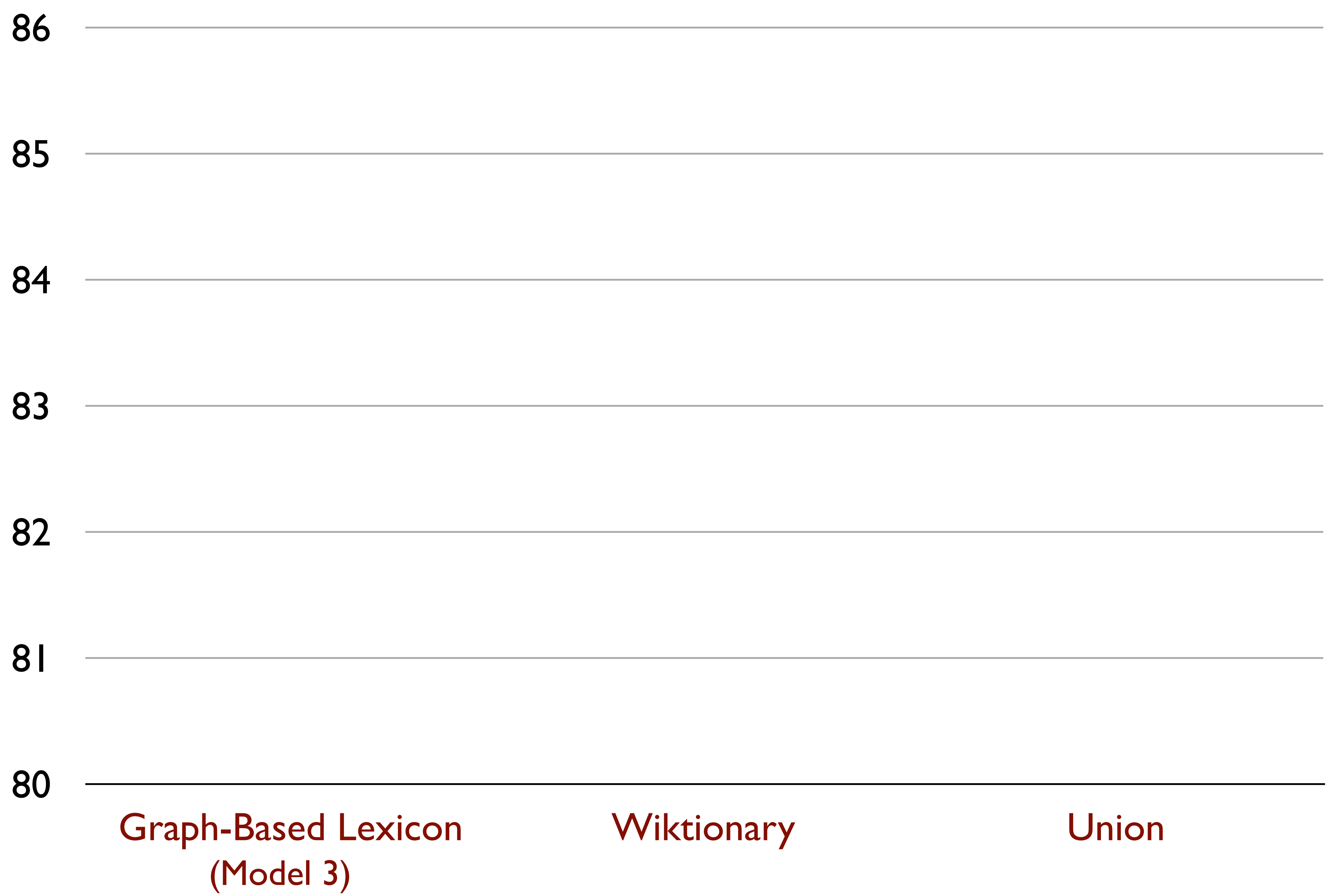
- Wiktionary
  - ~170 languages
  - Moving resource with regular snapshots

HMM with Wiktionary Constraints

Measured across 8 Indo-European Languages

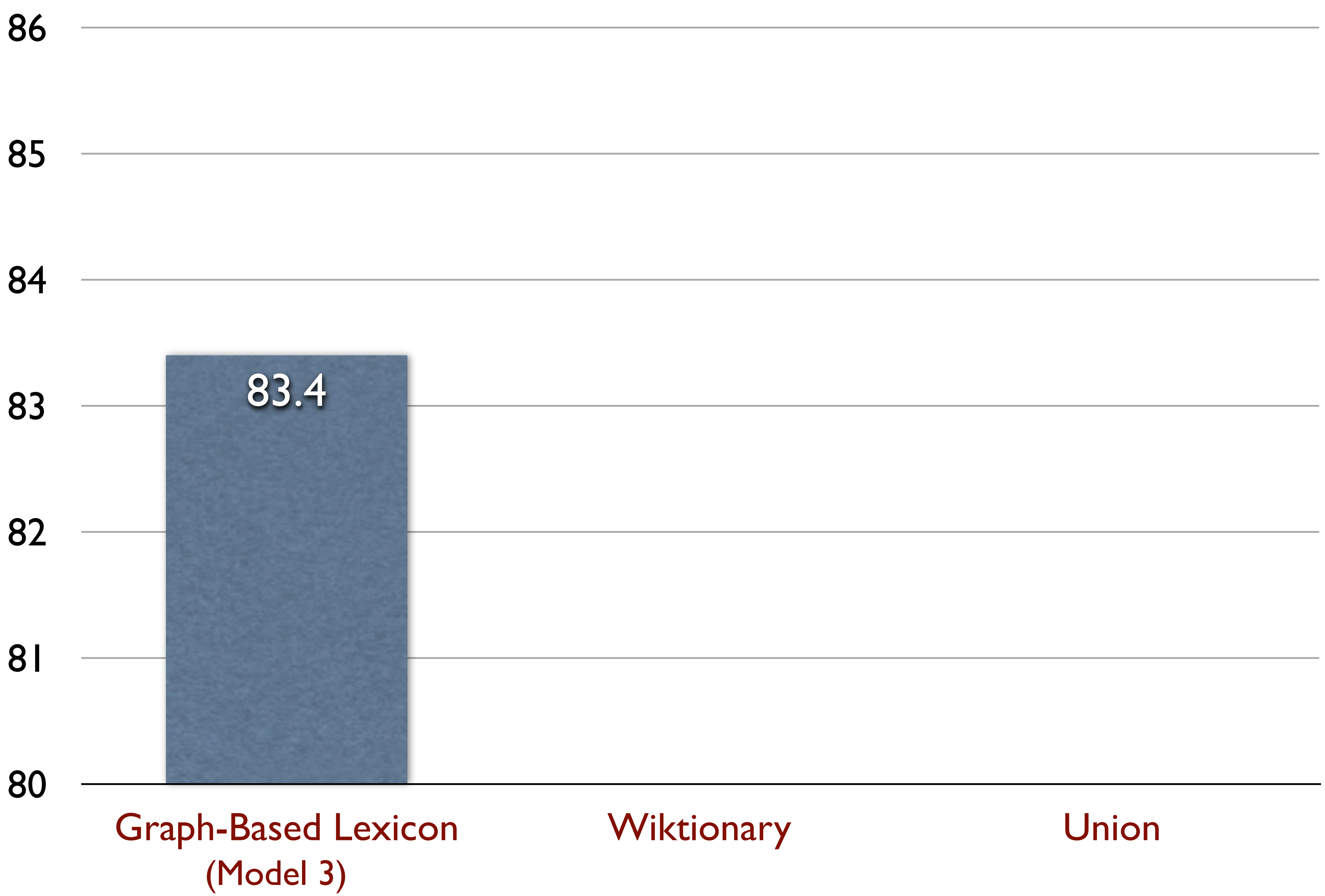
HMM with Wiktionary Constraints

Measured across 8 Indo-European Languages



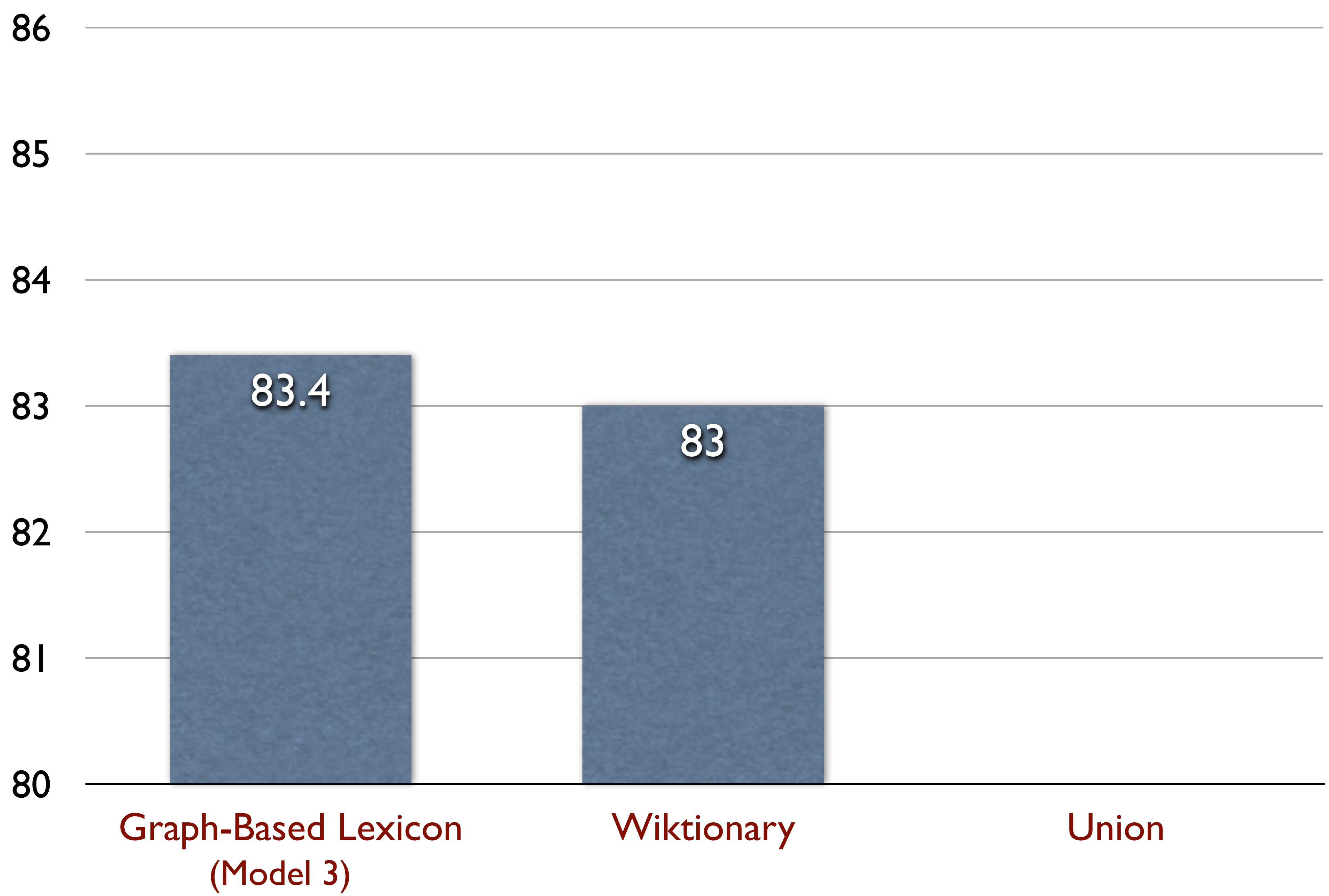
HMM with Wiktionary Constraints

Measured across 8 Indo-European Languages



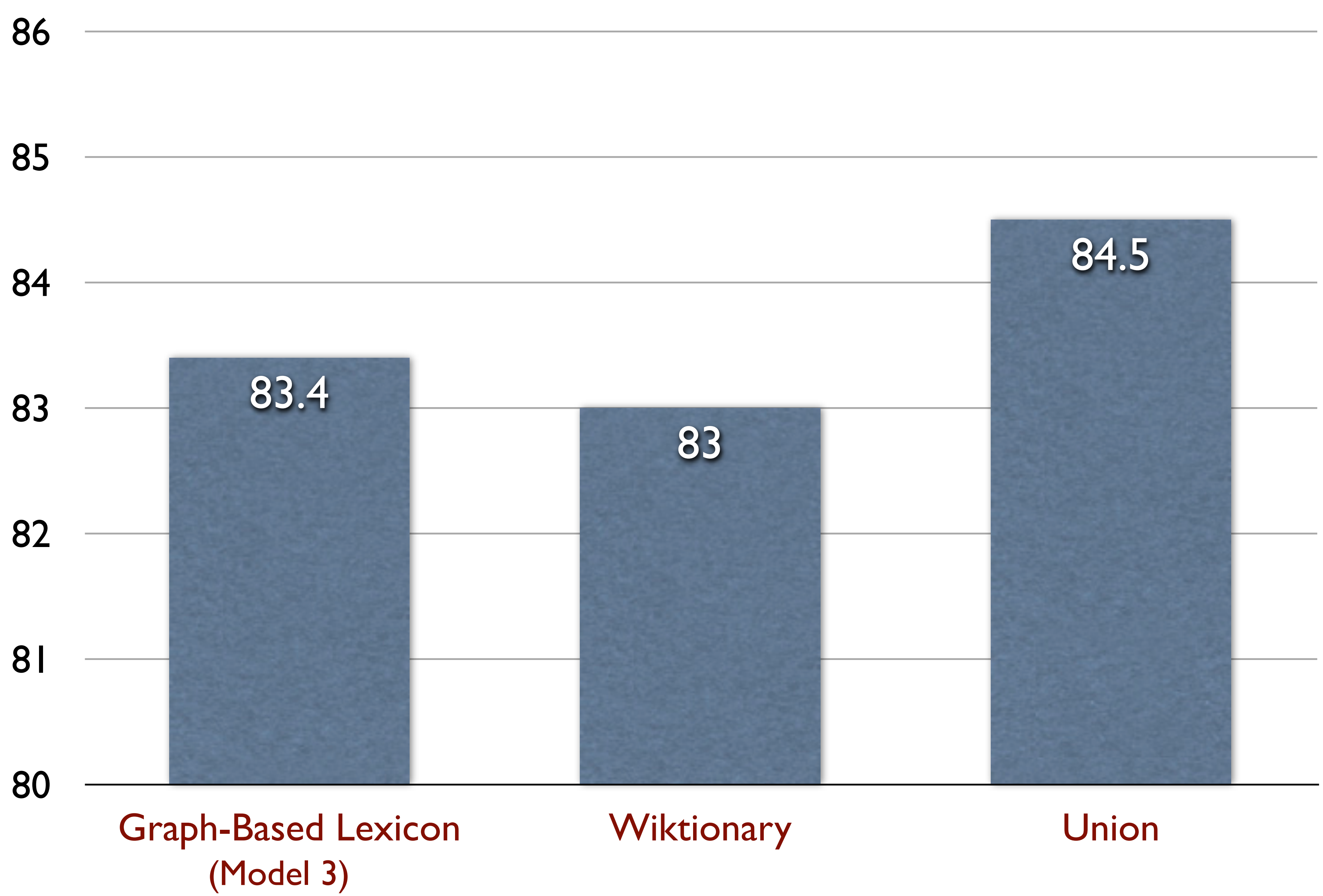
HMM with Wiktionary Constraints

Measured across 8 Indo-European Languages



HMM with Wiktionary Constraints

Measured across 8 Indo-European Languages



Complete **Token** Supervision

Produkterna måste vara helt rena

Complete **Token** Supervision

Produkterna	måste	vara	helt	rena
(The products)	(must)	(be)	(completely)	(pure)

Complete **Token** Supervision

Produkterna	måste	vara	helt	rena
(The products)	(must)	(be)	(completely)	(pure)
NOUN	VERB	VERB	ADV	ADJ

Complete **Token** Supervision

Produkterna	måste	vara	helt	rena
(The products)	(must)	(be)	(completely)	(pure)
NOUN	VERB	VERB	ADV	ADJ

Transition information in provided supervision

Complete **Token** Supervision

Produkterna	måste	vara	helt	rena
(The products)	(must)	(be)	(completely)	(pure)
NOUN	VERB	VERB	ADV	ADJ

Transition information in provided supervision

Discriminative model (Conditional Random Field):

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Complete **Token** Supervision

Produkterna	måste	vara	helt	rena
(The products)	(must)	(be)	(completely)	(pure)
NOUN	VERB	VERB	ADV	ADJ

Transition information in provided supervision

Discriminative model (Conditional Random Field):

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}')}$$

All possible tag sequences for $\boldsymbol{x}$

Parallel Sentence Pair

Produkterna måste vara helt rena

Parallel Sentence Pair

The agricultural products must be pure

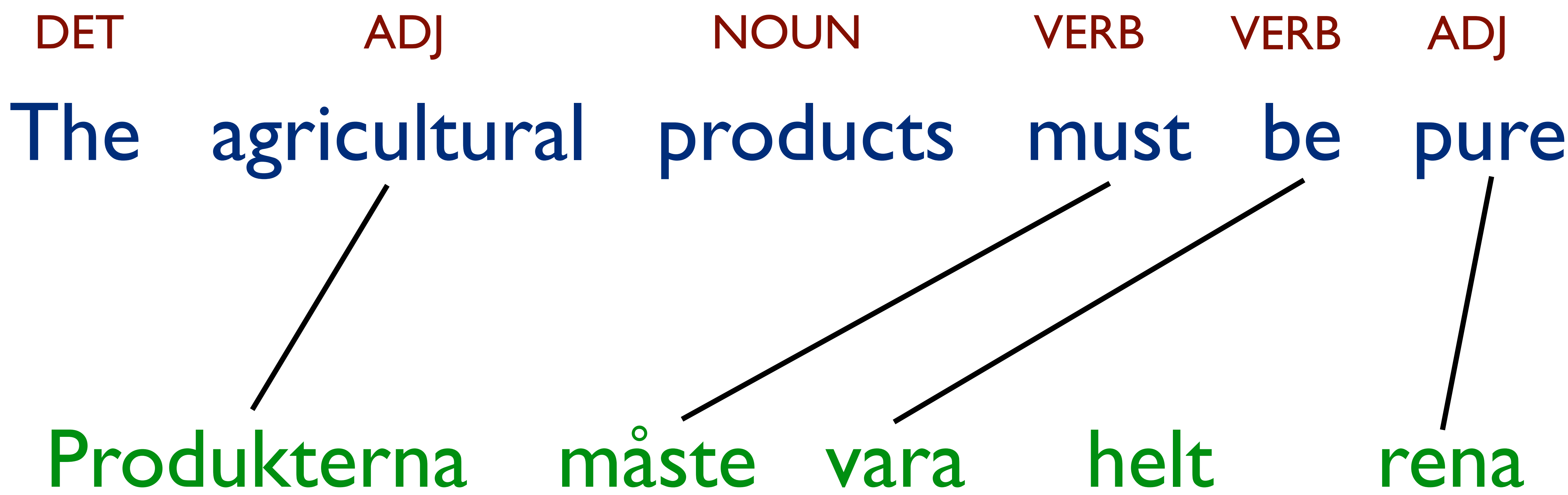
Produkterna måste vara helt rena

Parallel Sentence Pair

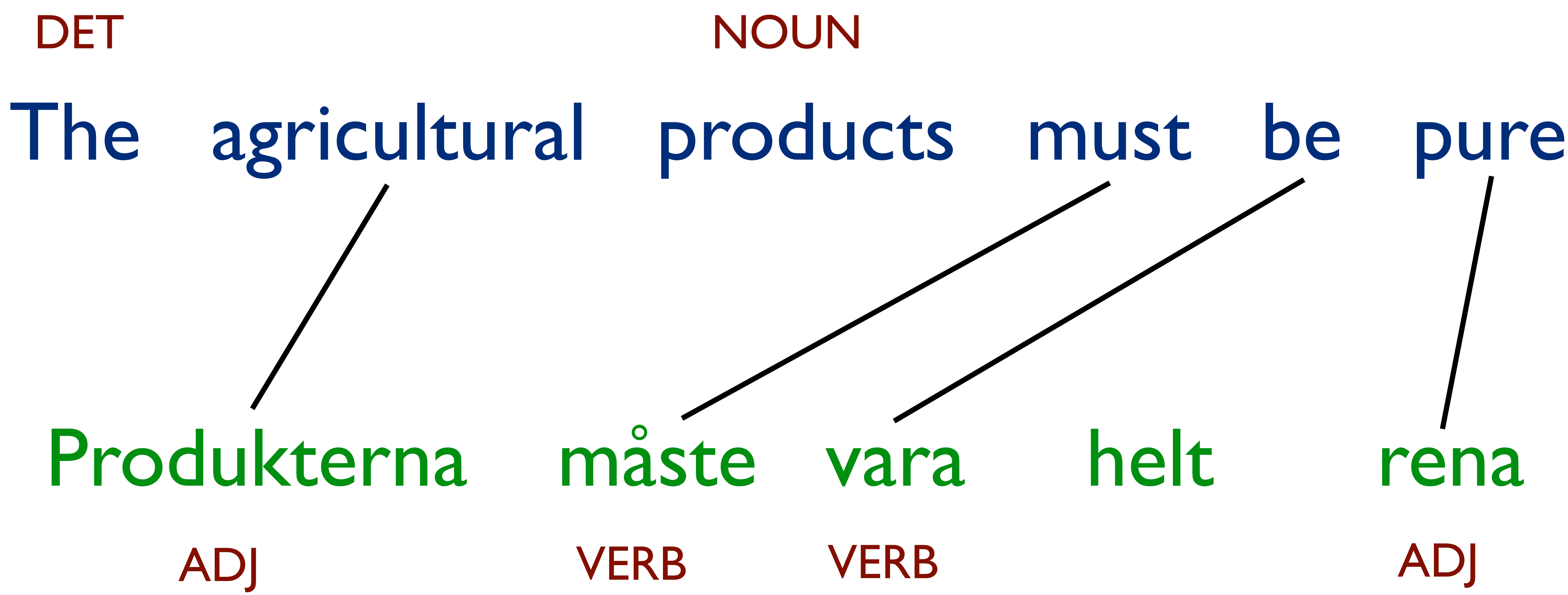
DET	ADJ	NOUN	VERB	VERB	ADJ
The	agricultural	products	must	be	pure

Produkterna måste vara helt rena

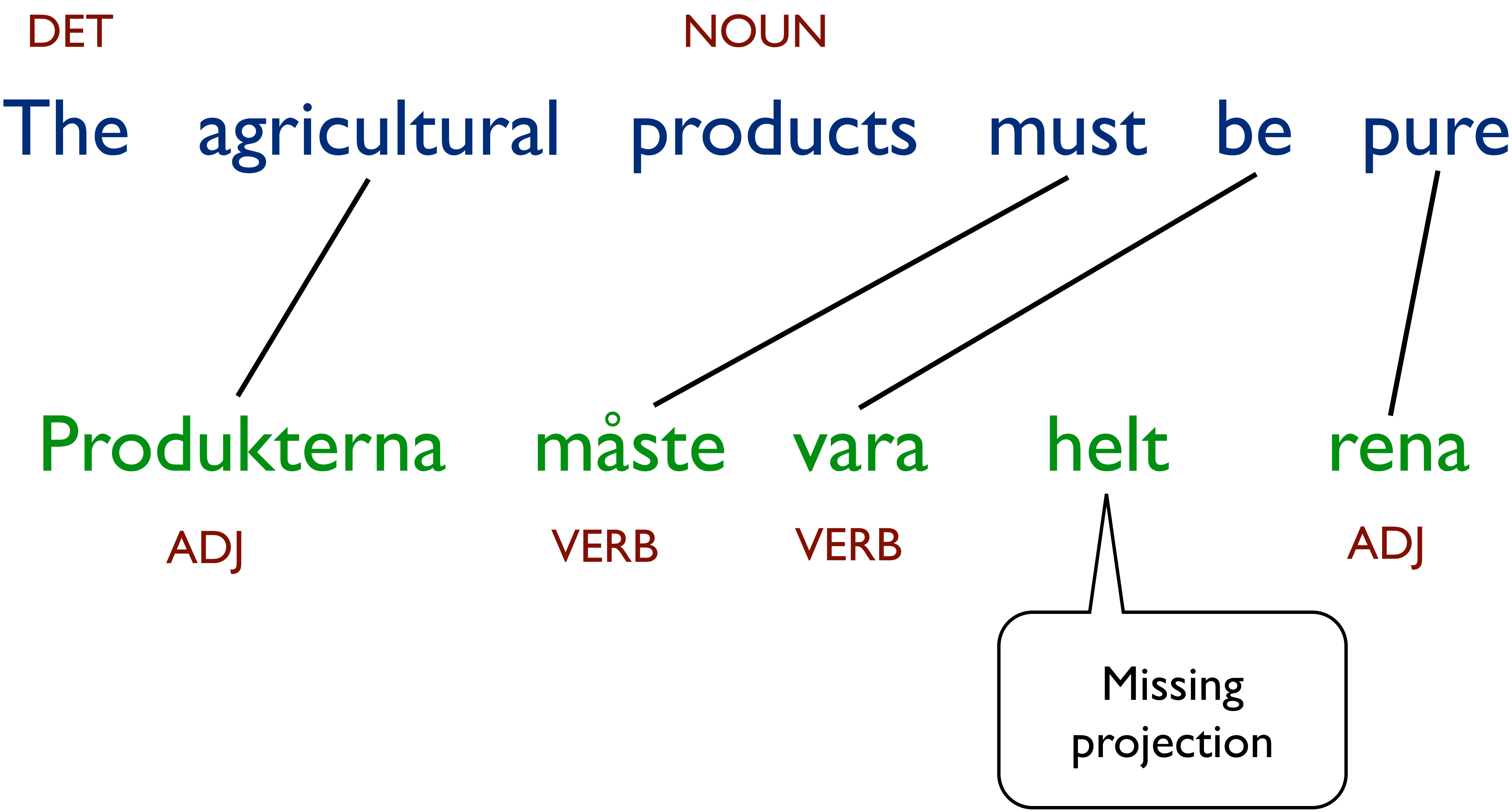
Parallel Sentence Pair



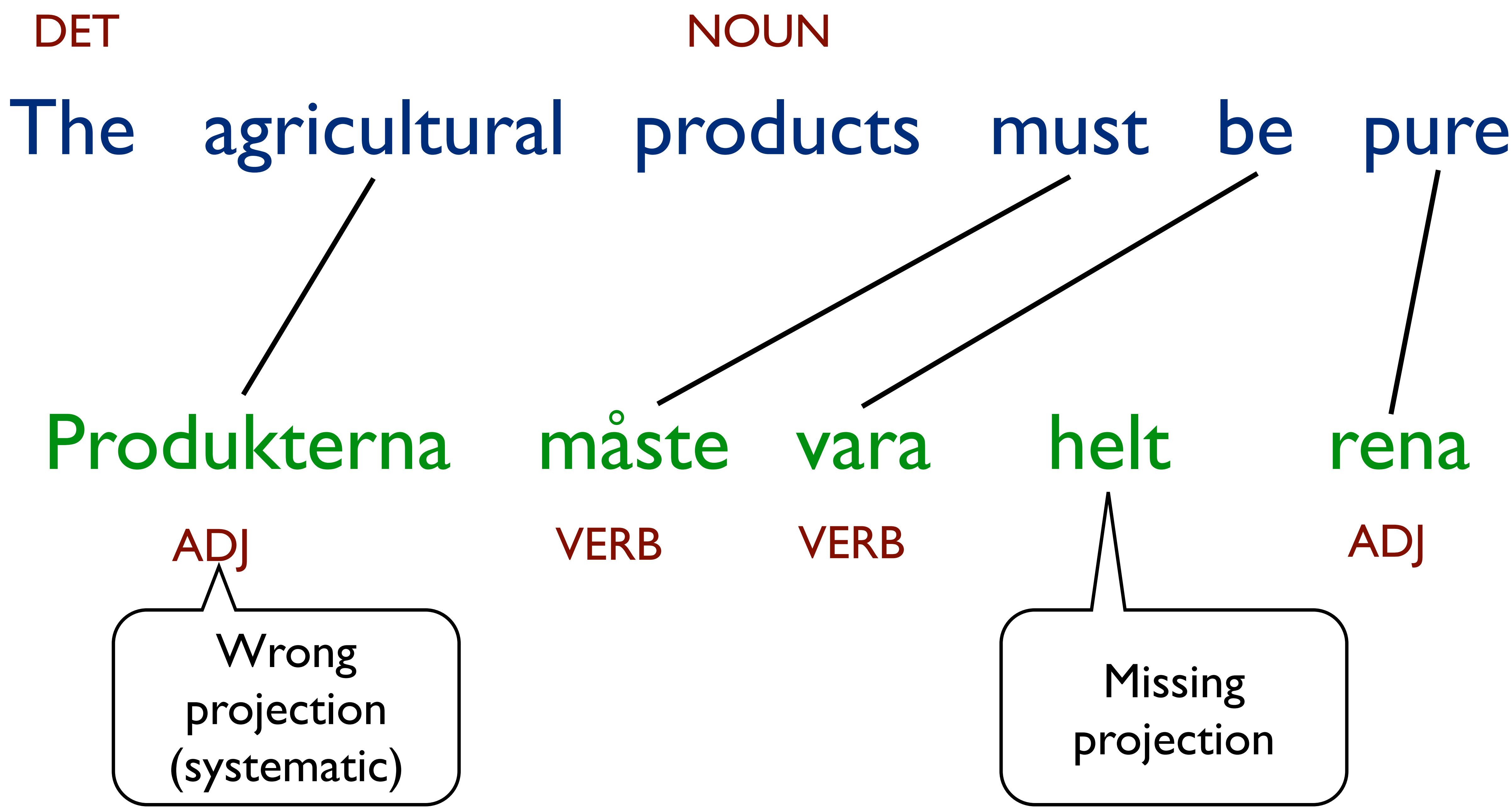
Parallel Sentence Pair



Parallel Sentence Pair



Parallel Sentence Pair



Incomplete **Token** Level Supervision

Produkterna	måste	vara	helt	rena
ADJ	VERB	VERB		ADJ

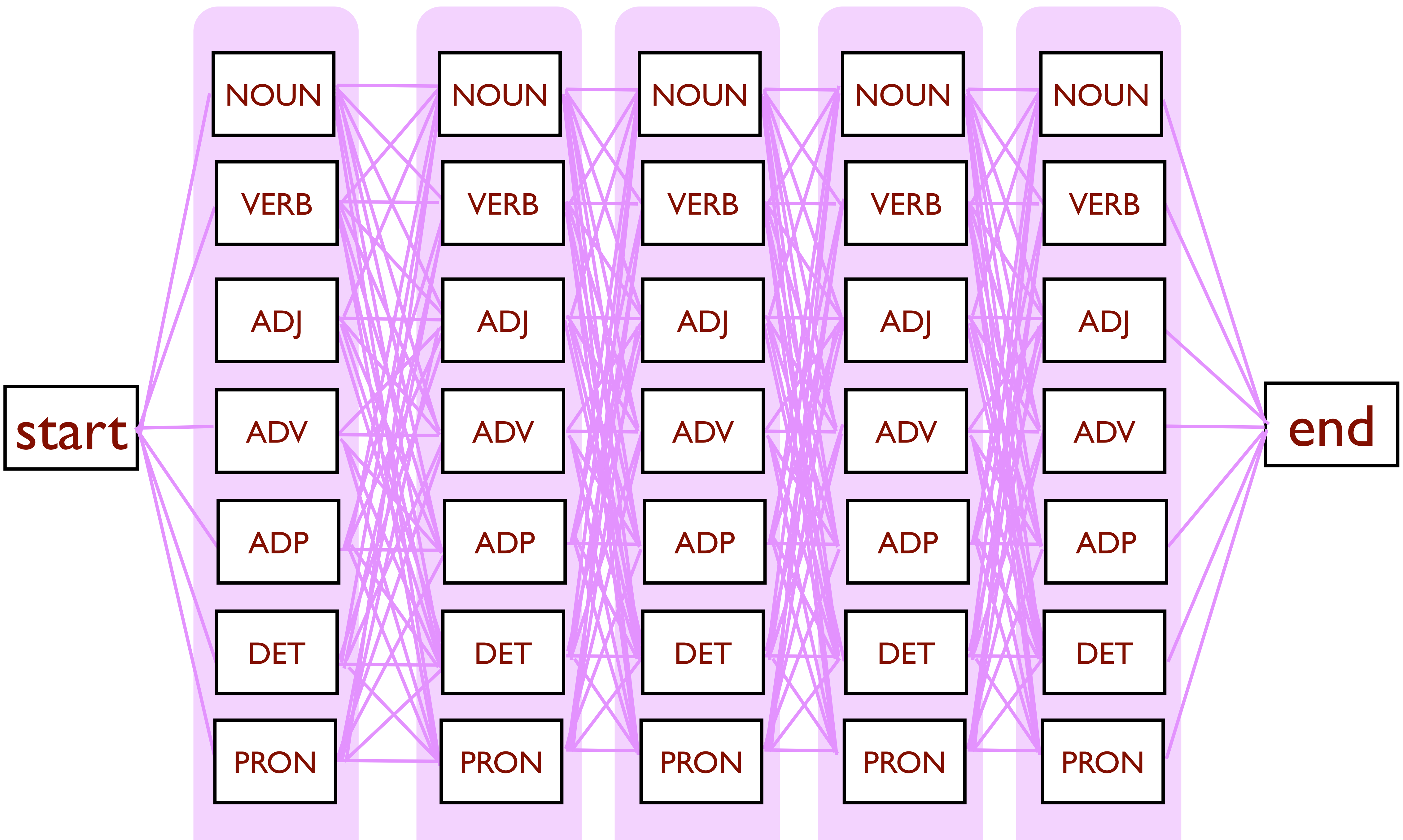
Incomplete **Token** Level Supervision

Produkterna	måste	vara	helt	rena
ADJ	VERB	VERB		ADJ

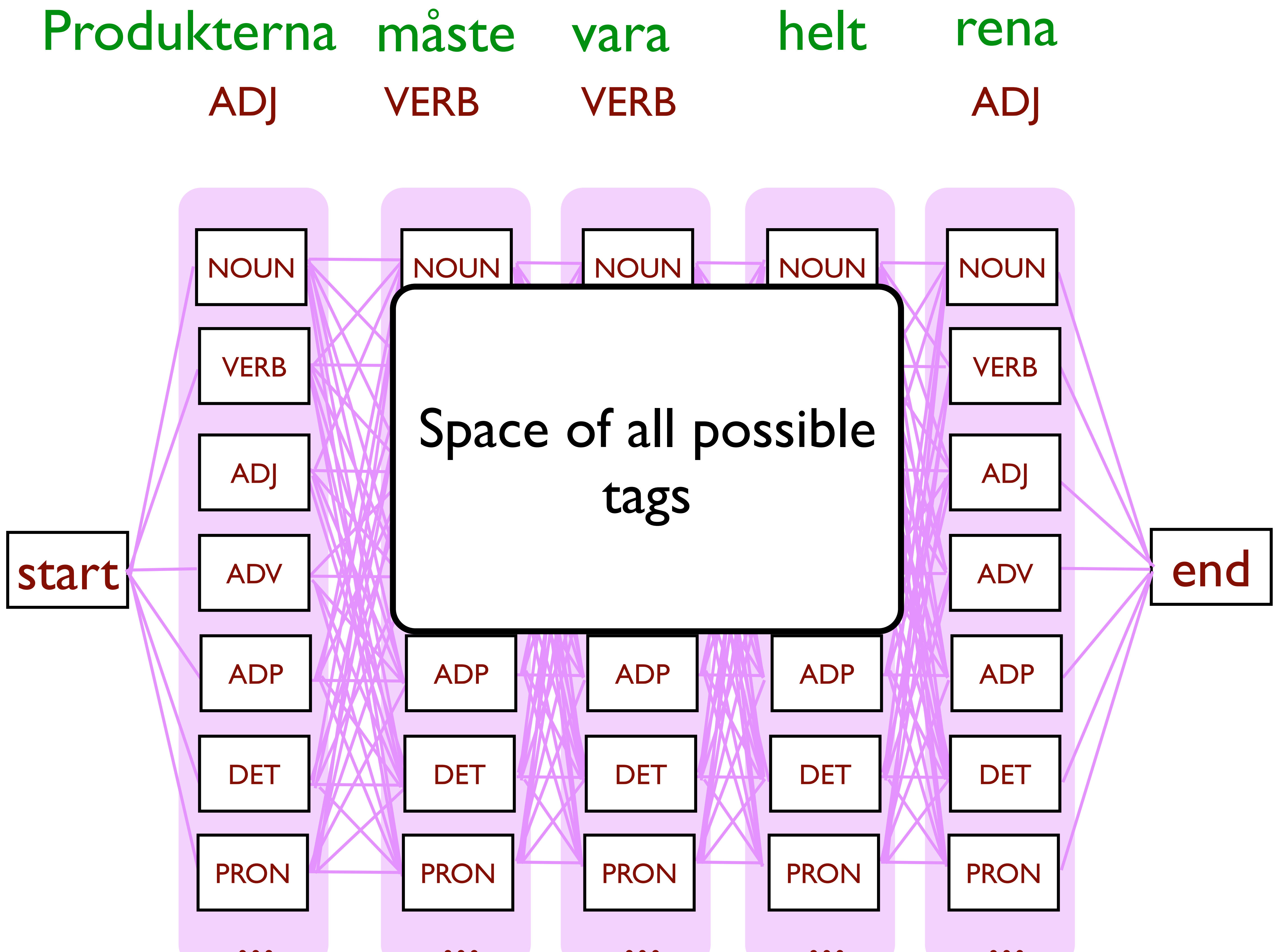
projected tags

Incomplete **Token** Level Supervision

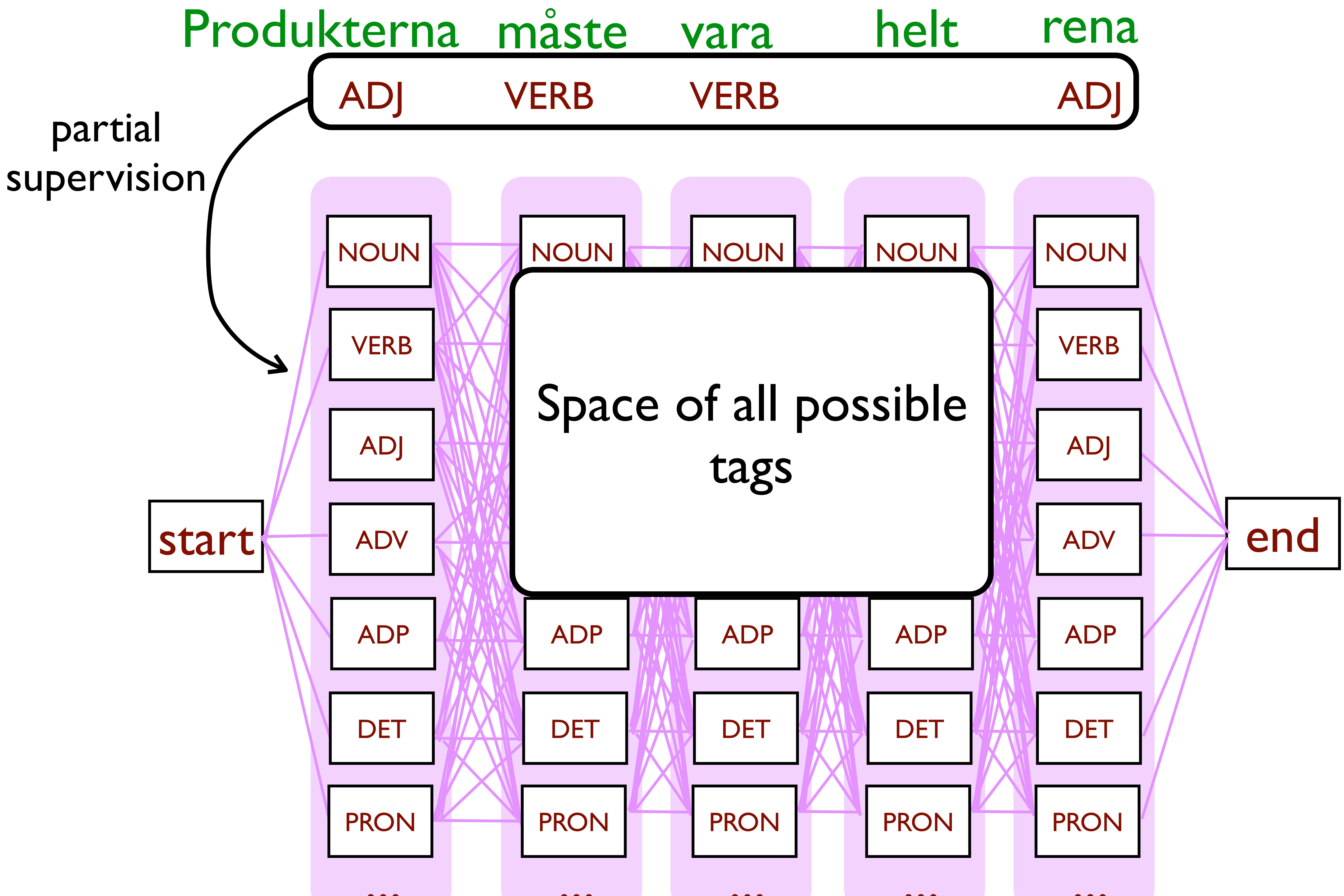
Produkterna	måste	vara	helt	rena
ADJ	VERB	VERB		ADJ



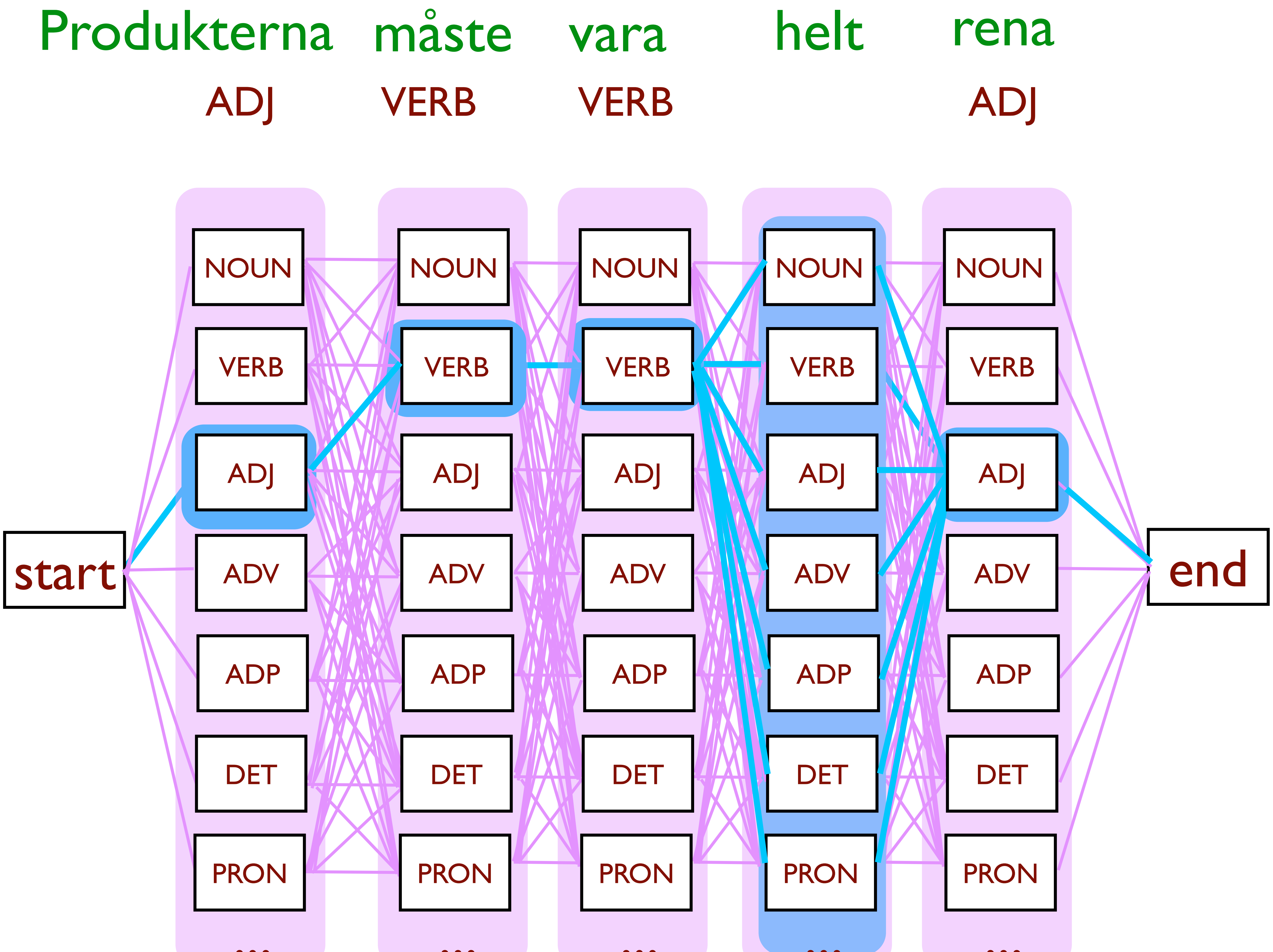
Incomplete **Token** Level Supervision



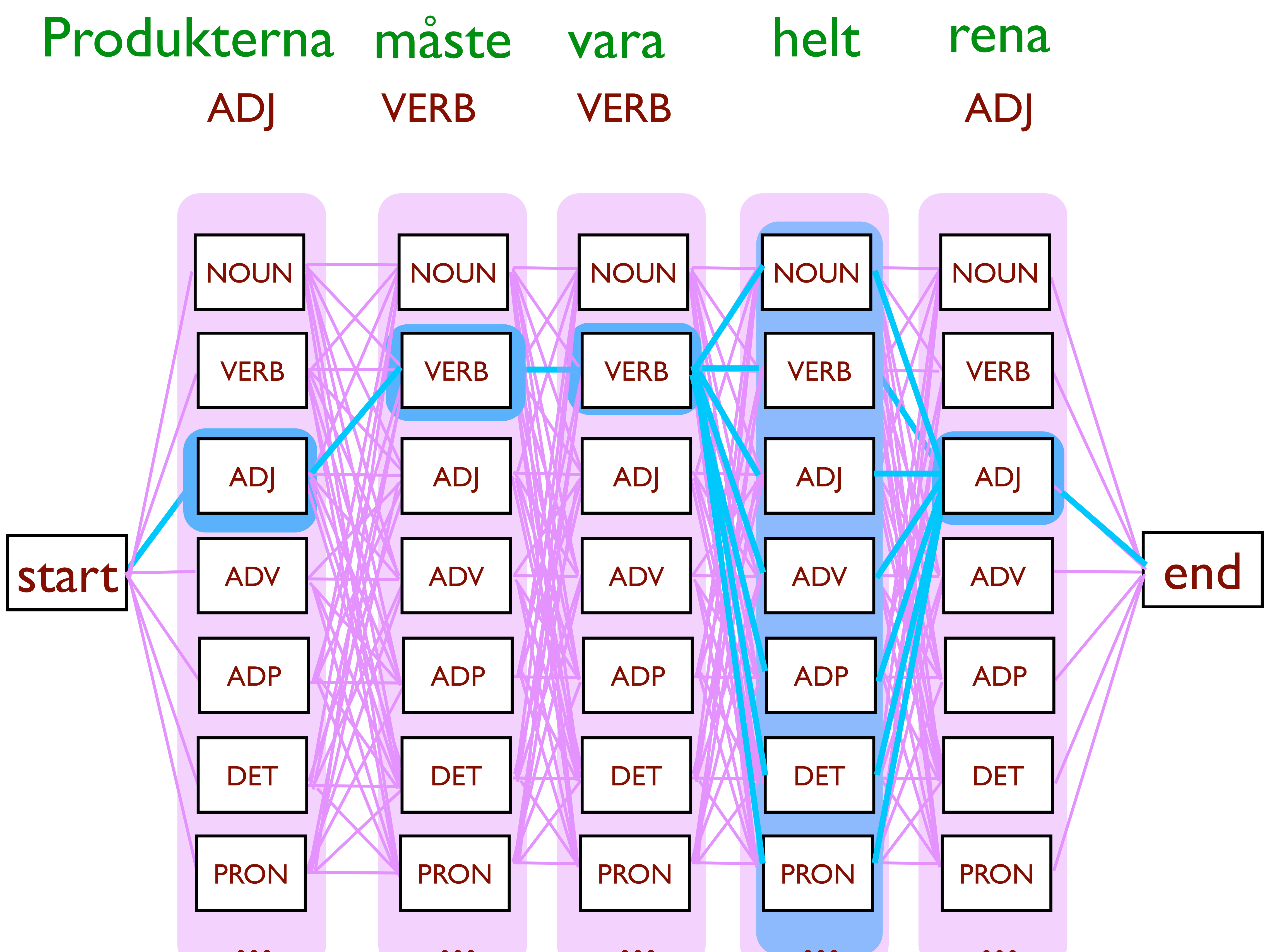
Incomplete **Token** Level Supervision



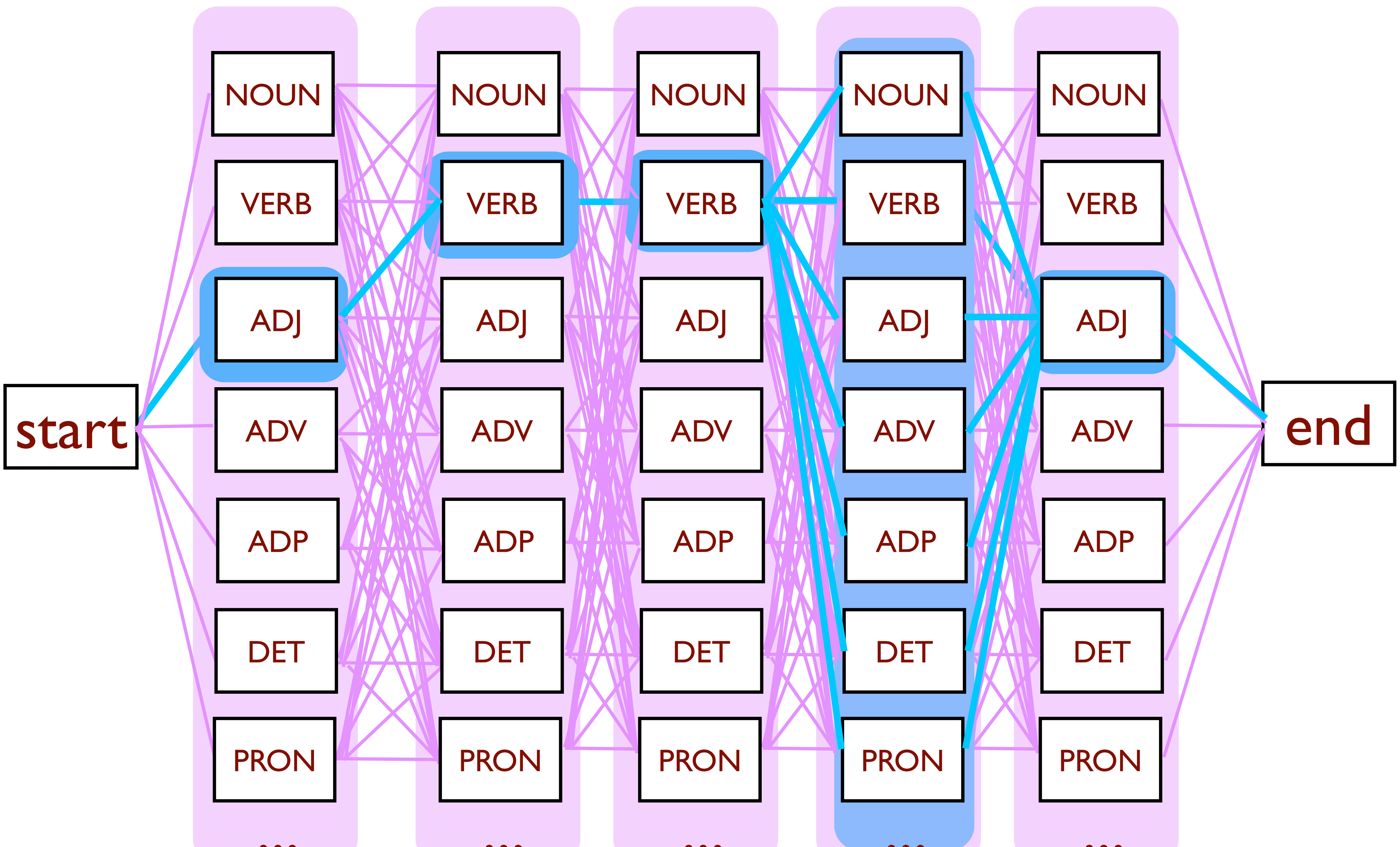
Incomplete **Token** Level Supervision



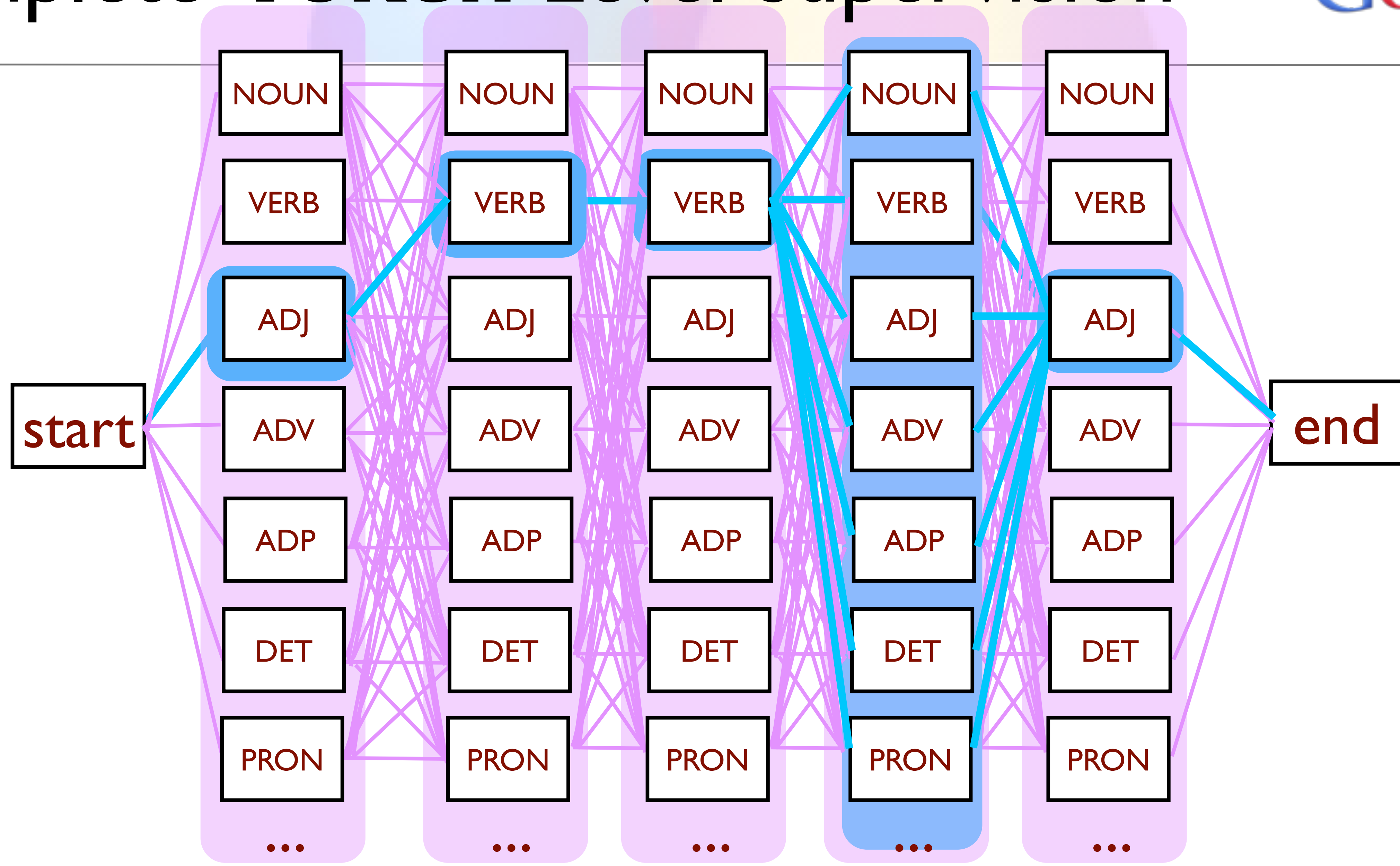
Incomplete **Token** Level Supervision



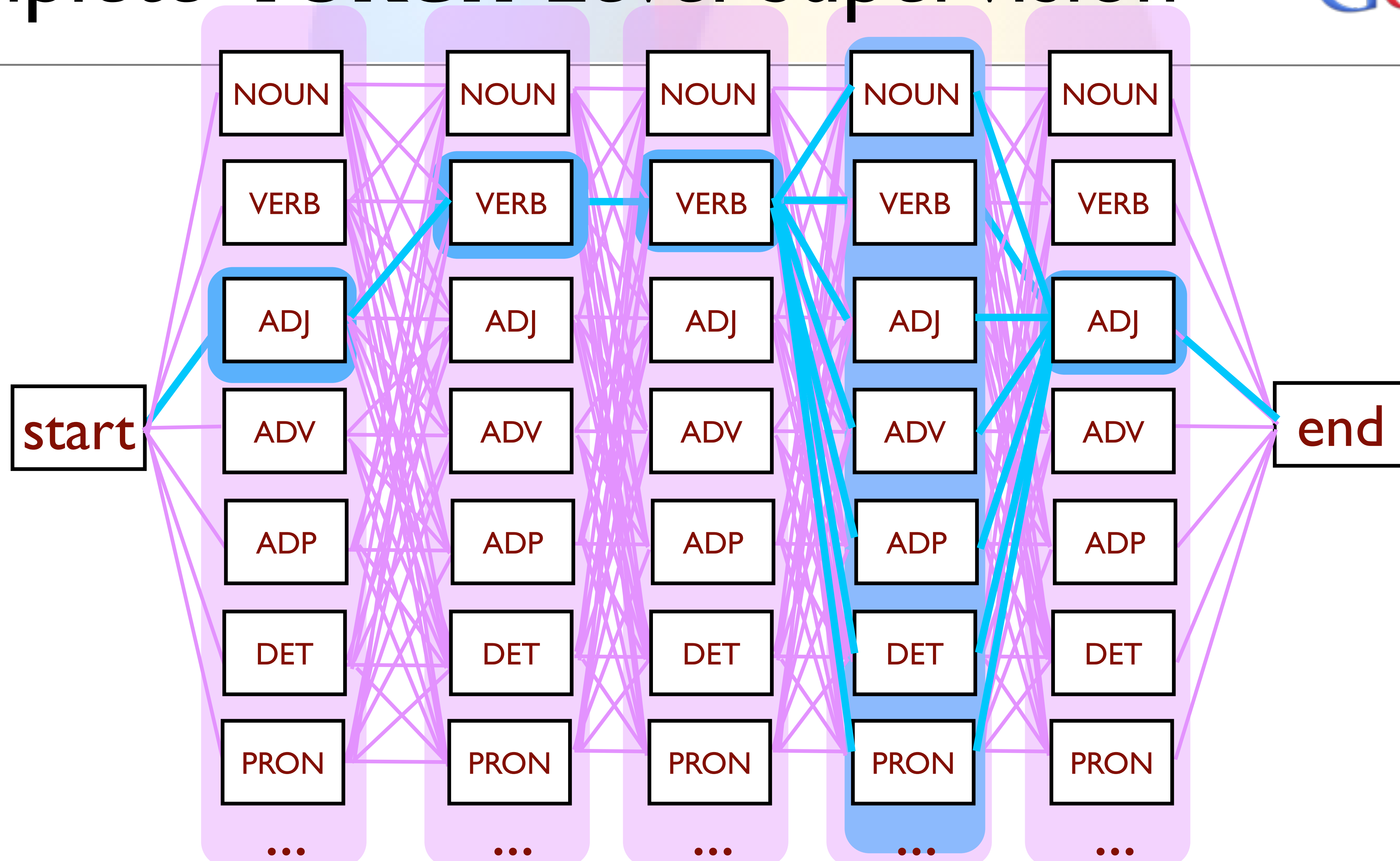
Incomplete **Token** Level Supervision



Incomplete **Token** Level Supervision

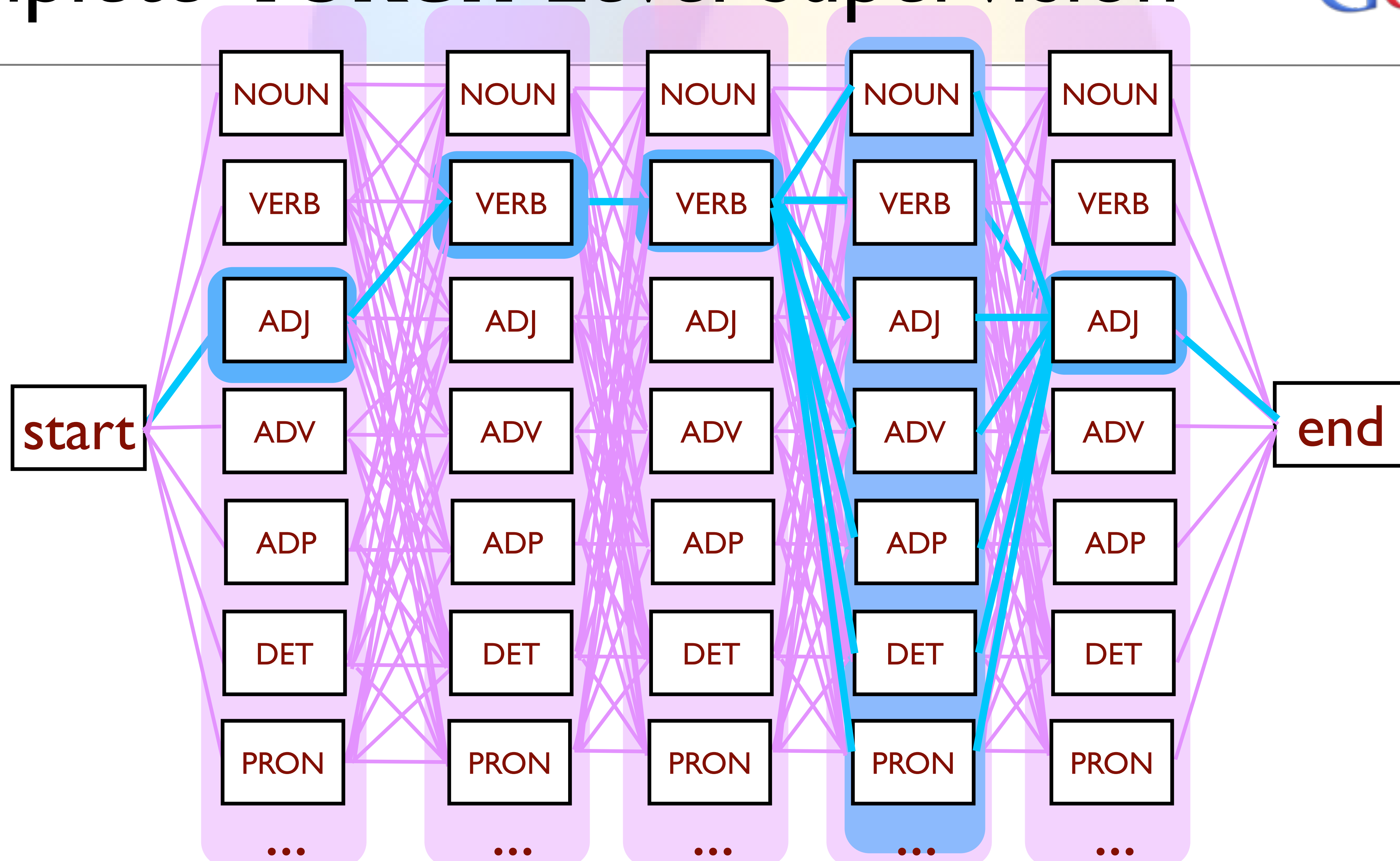


Incomplete **Token** Level Supervision



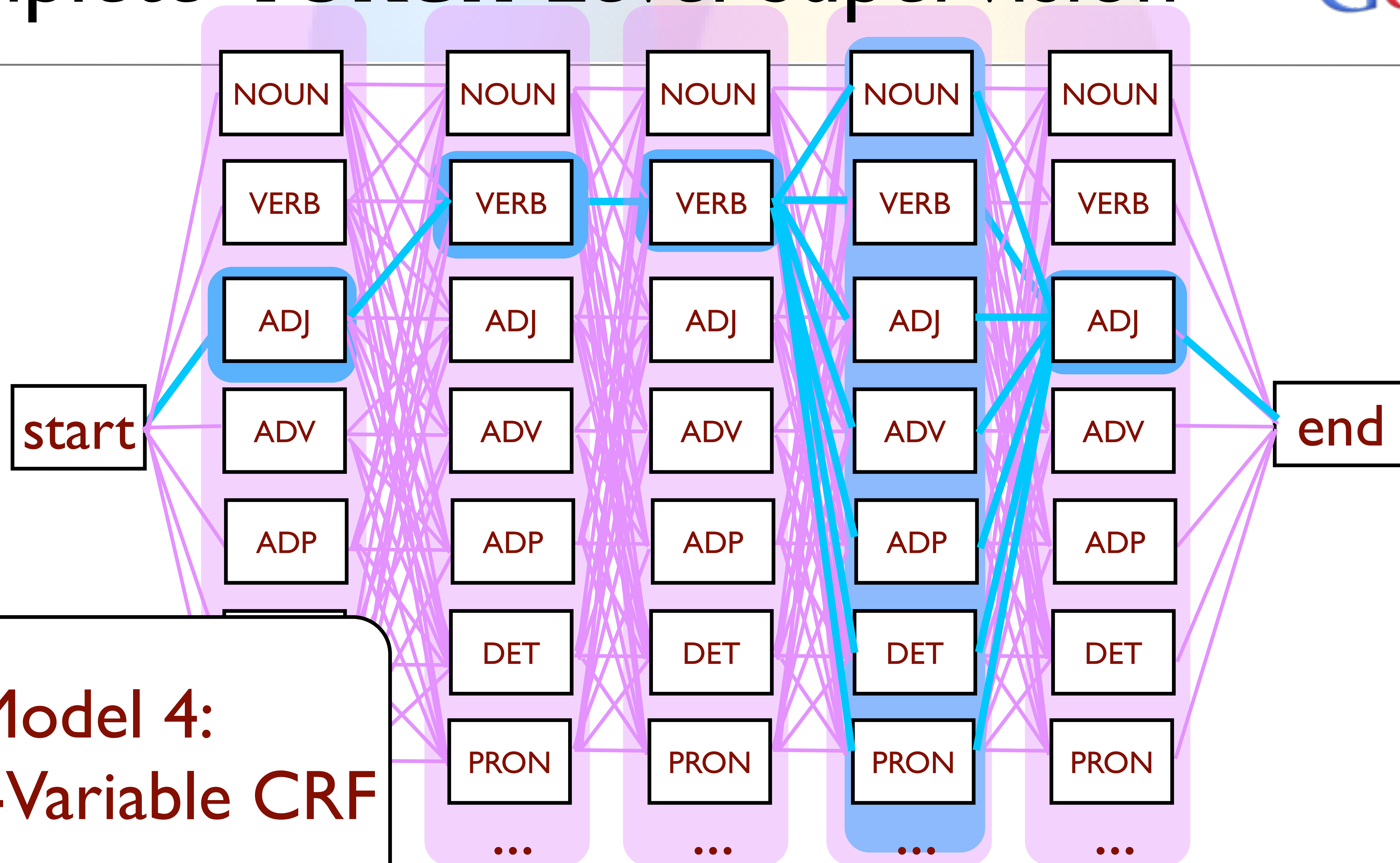
$$p(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = \frac{\displaystyle\sum_{\boldsymbol{y}\in\tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\displaystyle\sum_{\boldsymbol{y}'\in\mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Incomplete **Token** Level Supervision



$$p(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = \frac{\sum_{\boldsymbol{y} \in \tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Incomplete **Token** Level Supervision

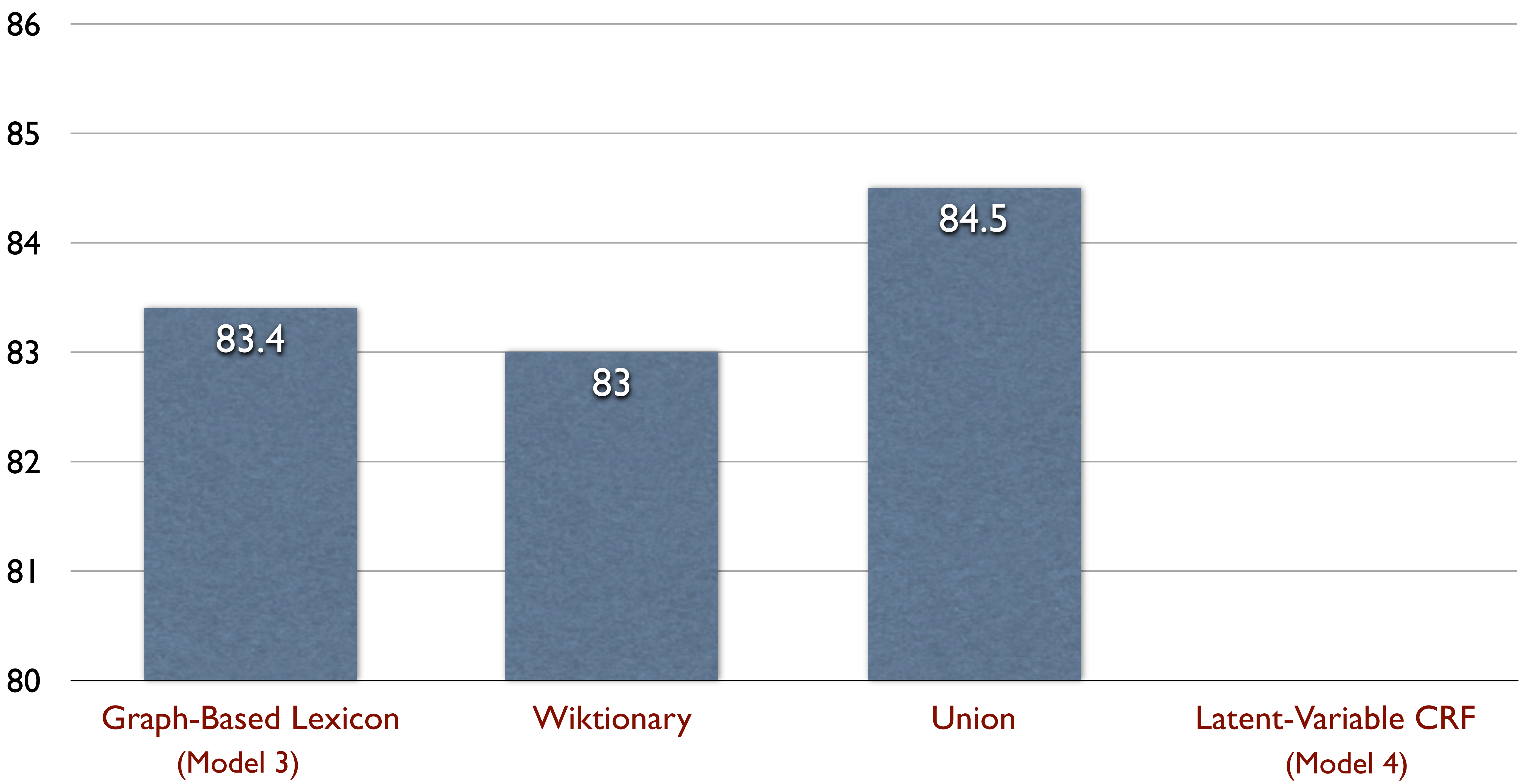


Model 4:
Latent-Variable CRF

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = \frac{\sum_{\boldsymbol{y} \in \tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Model Comparison

Measured across 8 Indo-European Languages

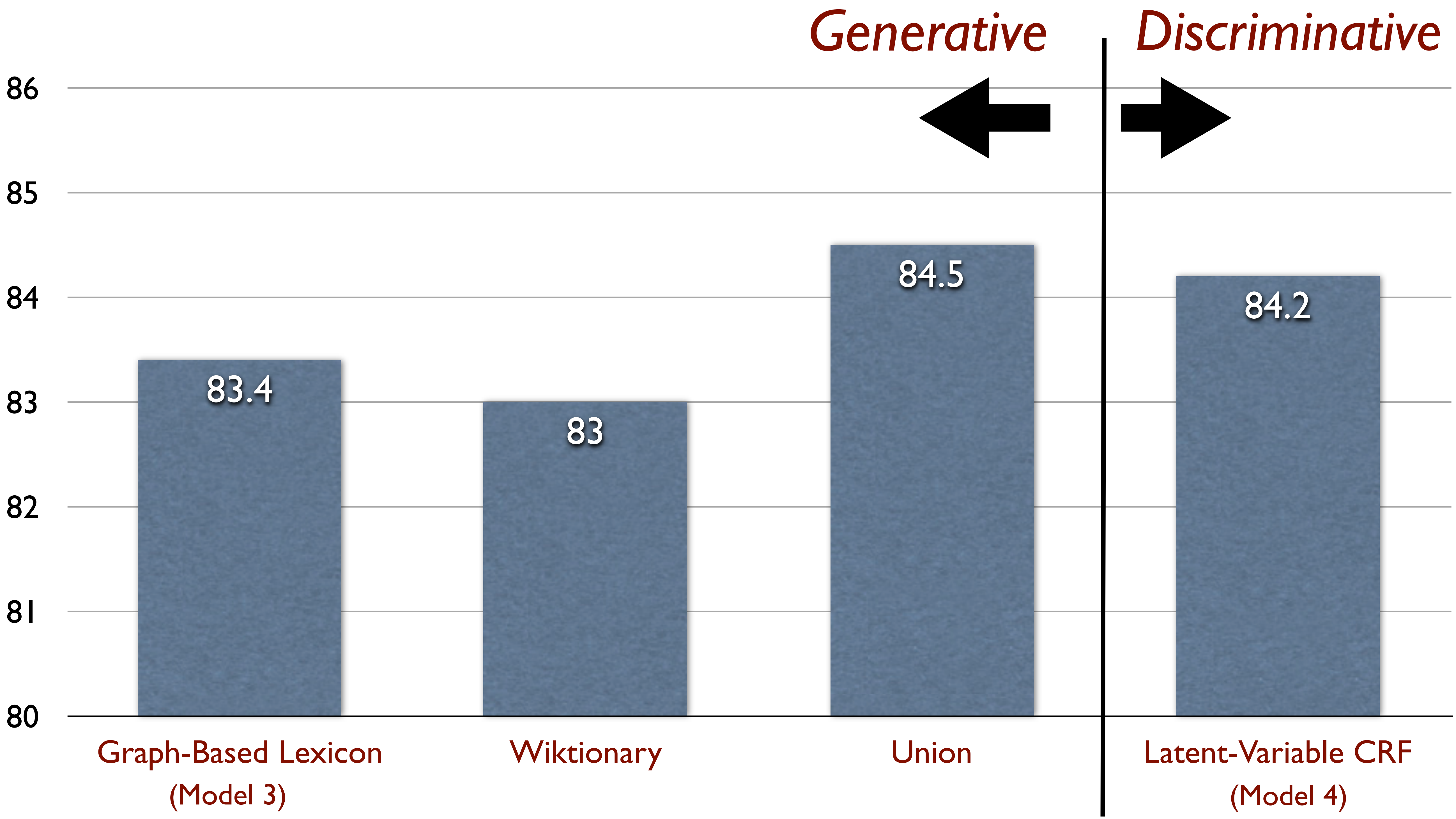


Model Comparison

Measured across 8 Indo-European Languages

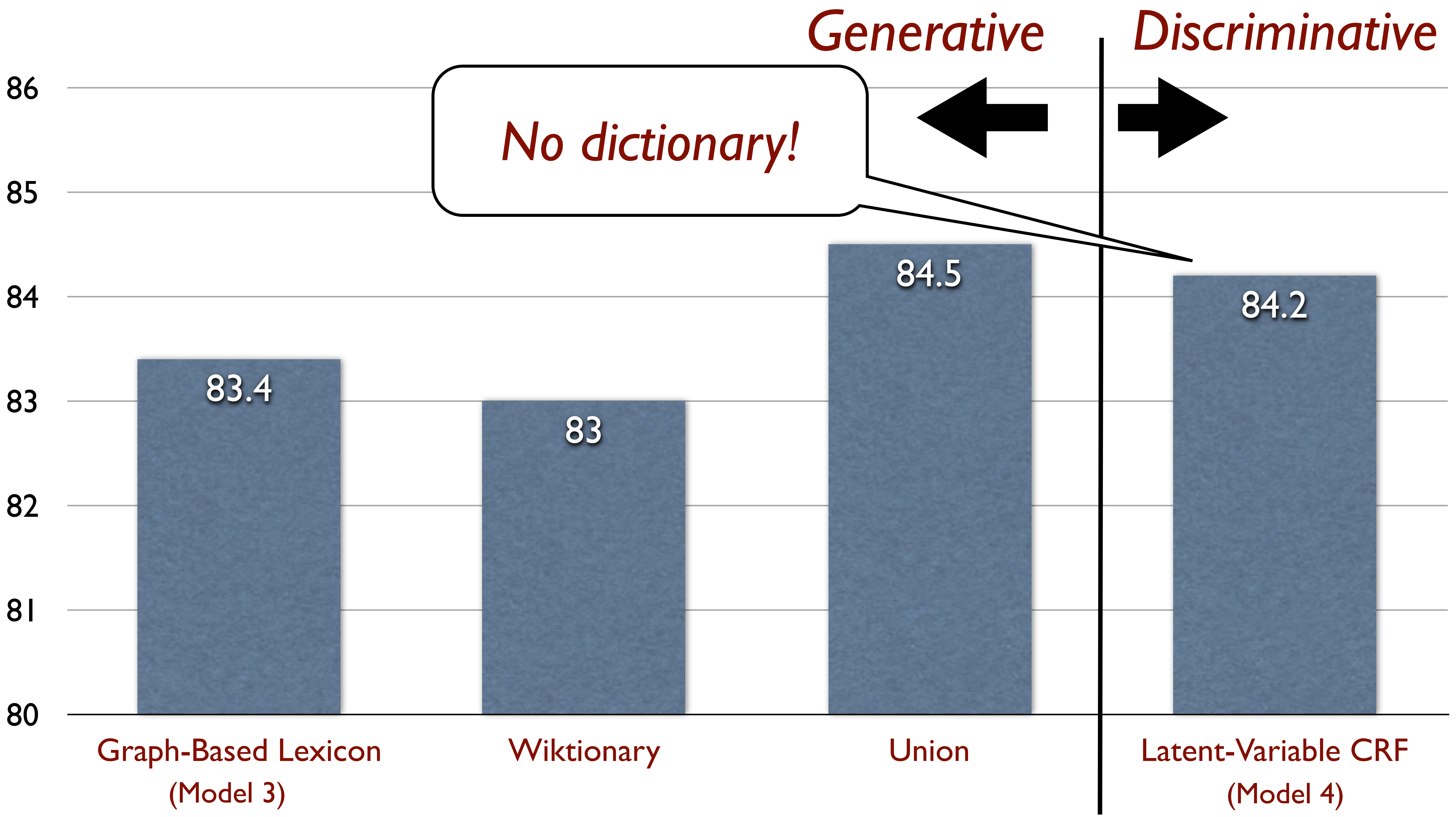
Model Comparison

Measured across 8 Indo-European Languages

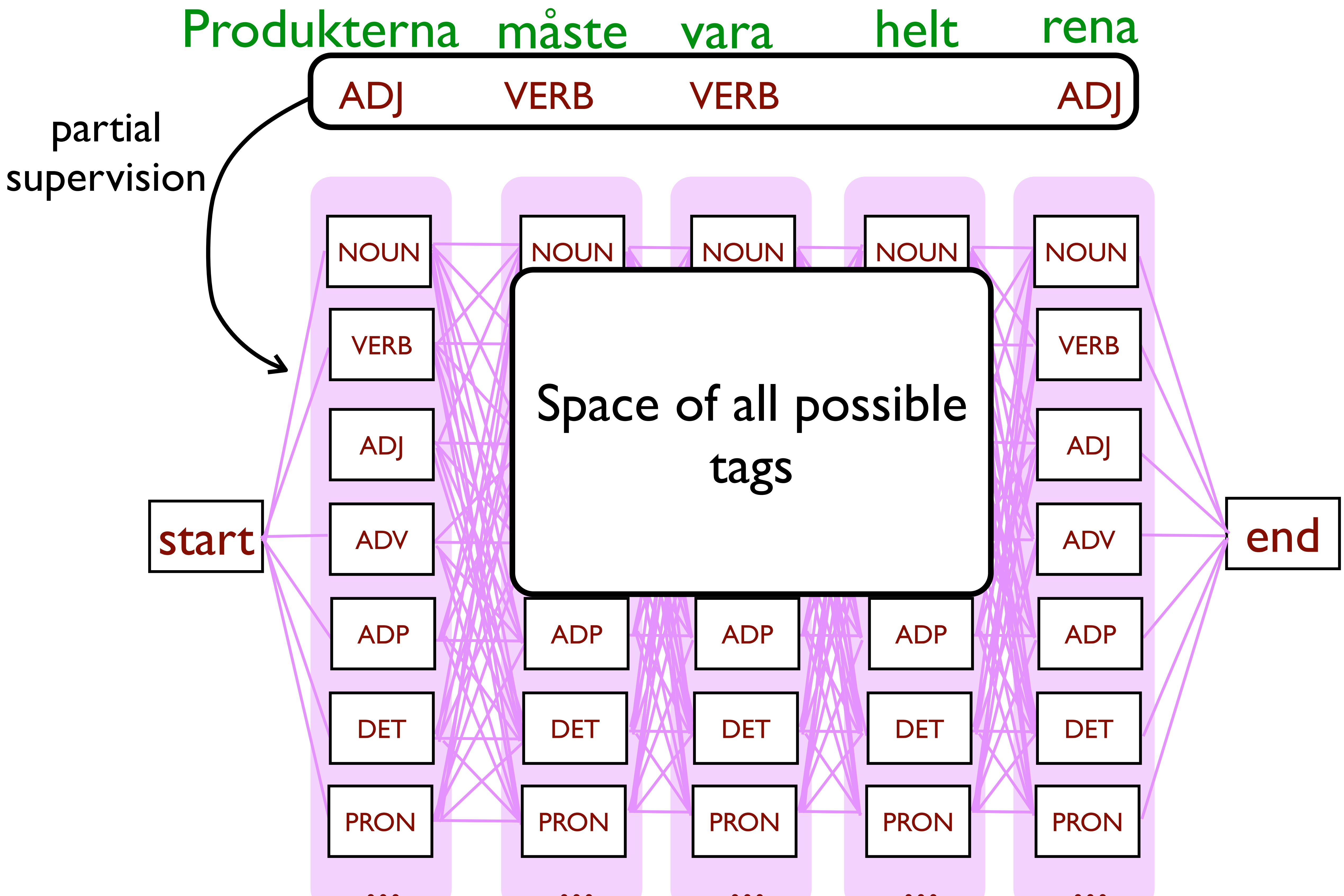


Model Comparison

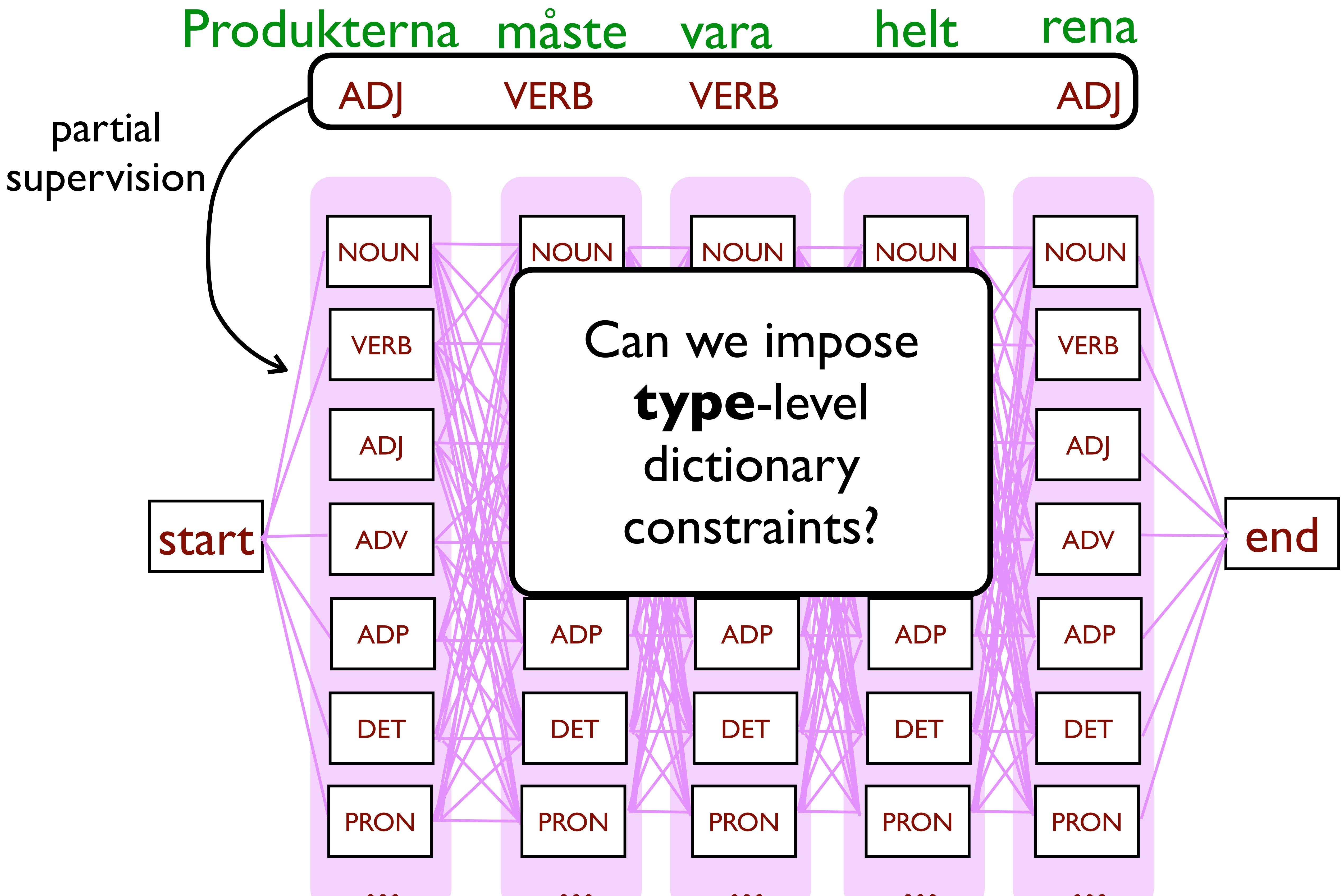
Measured across 8 Indo-European Languages



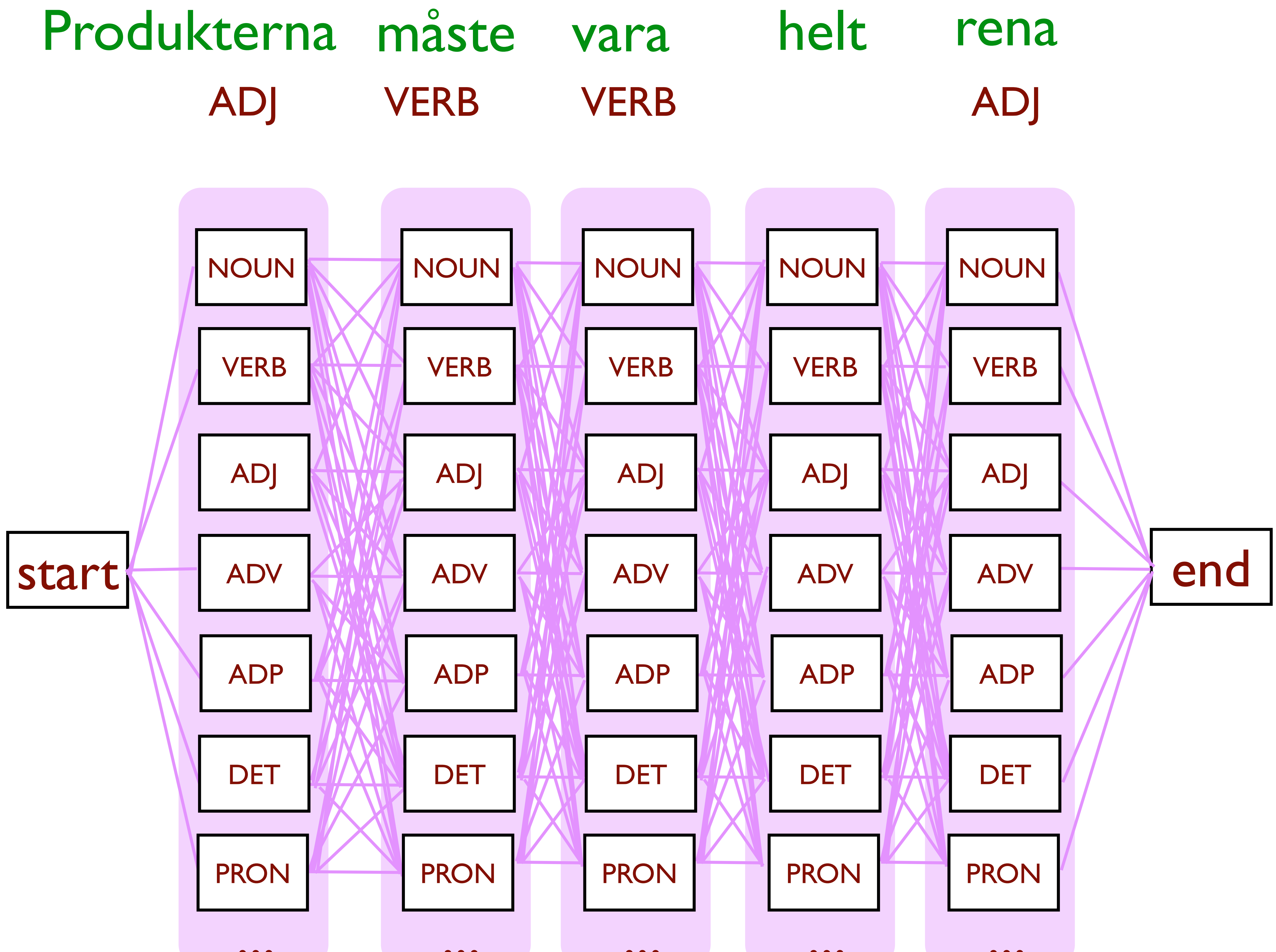
Coupled **Token** and **Type** Constraints Google



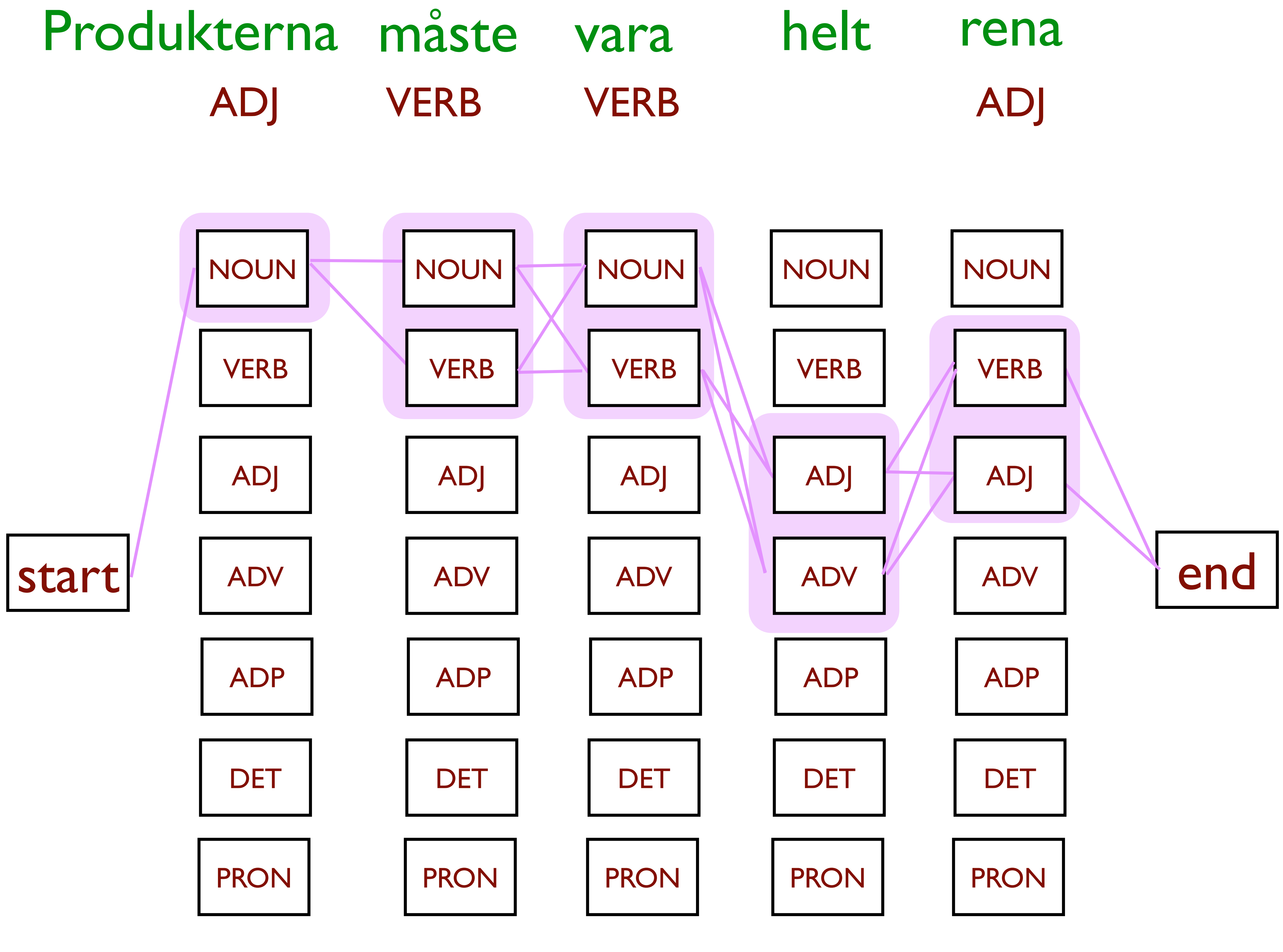
Coupled **Token** and **Type** Constraints



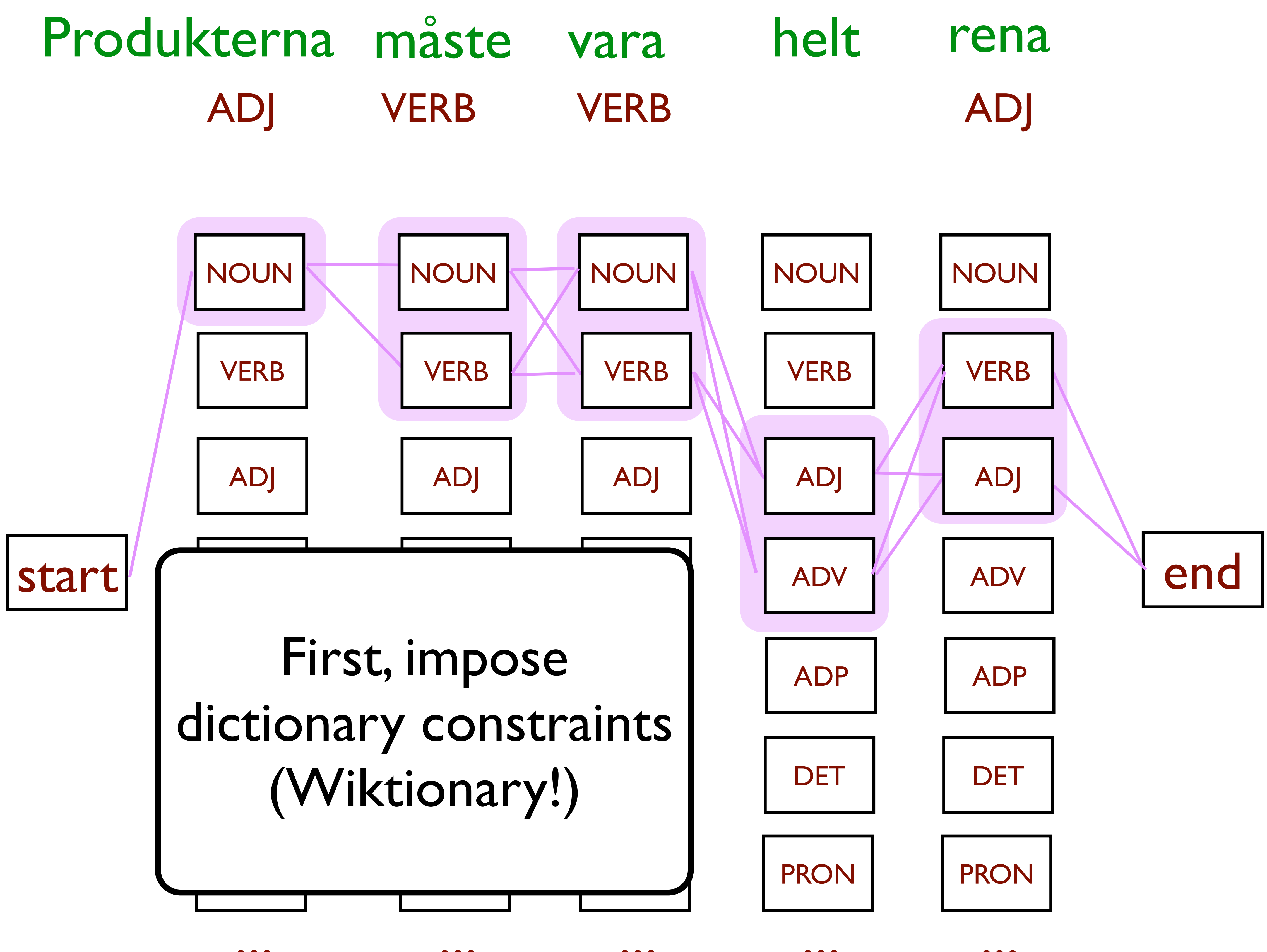
Coupled **Token** and **Type** Constraints Google



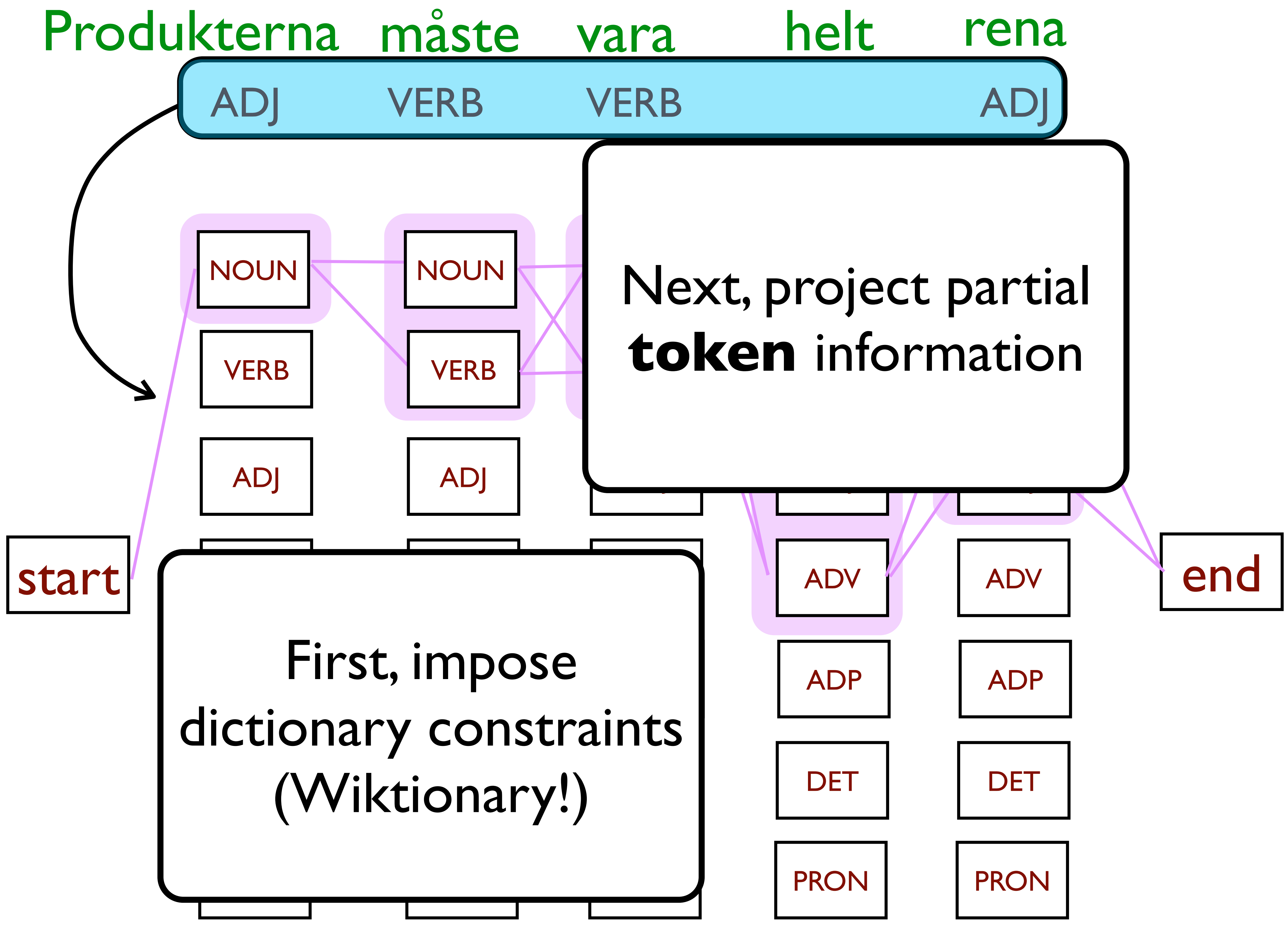
Coupled **Token** and **Type** Constraints



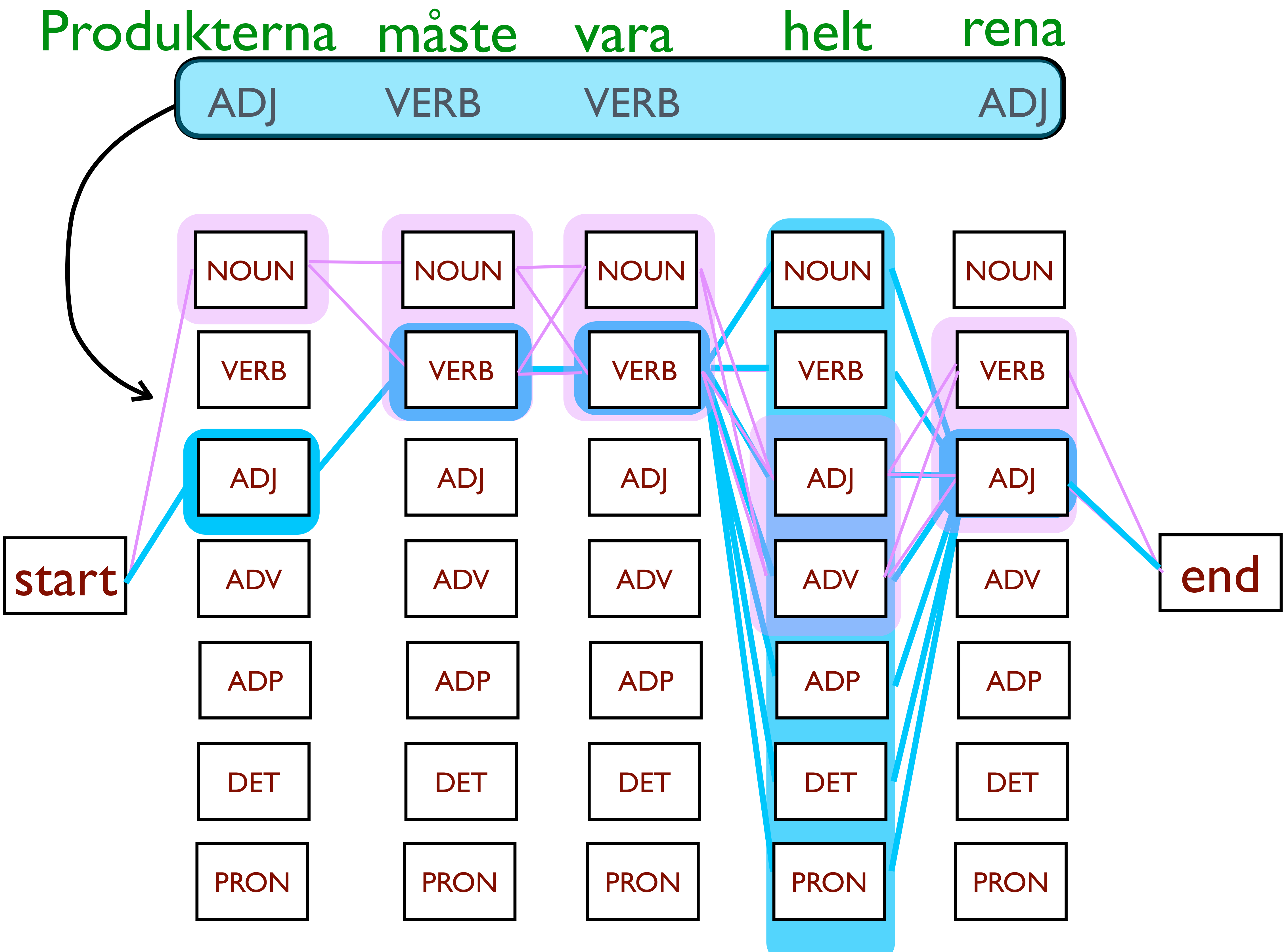
Coupled **Token** and **Type** Constraints



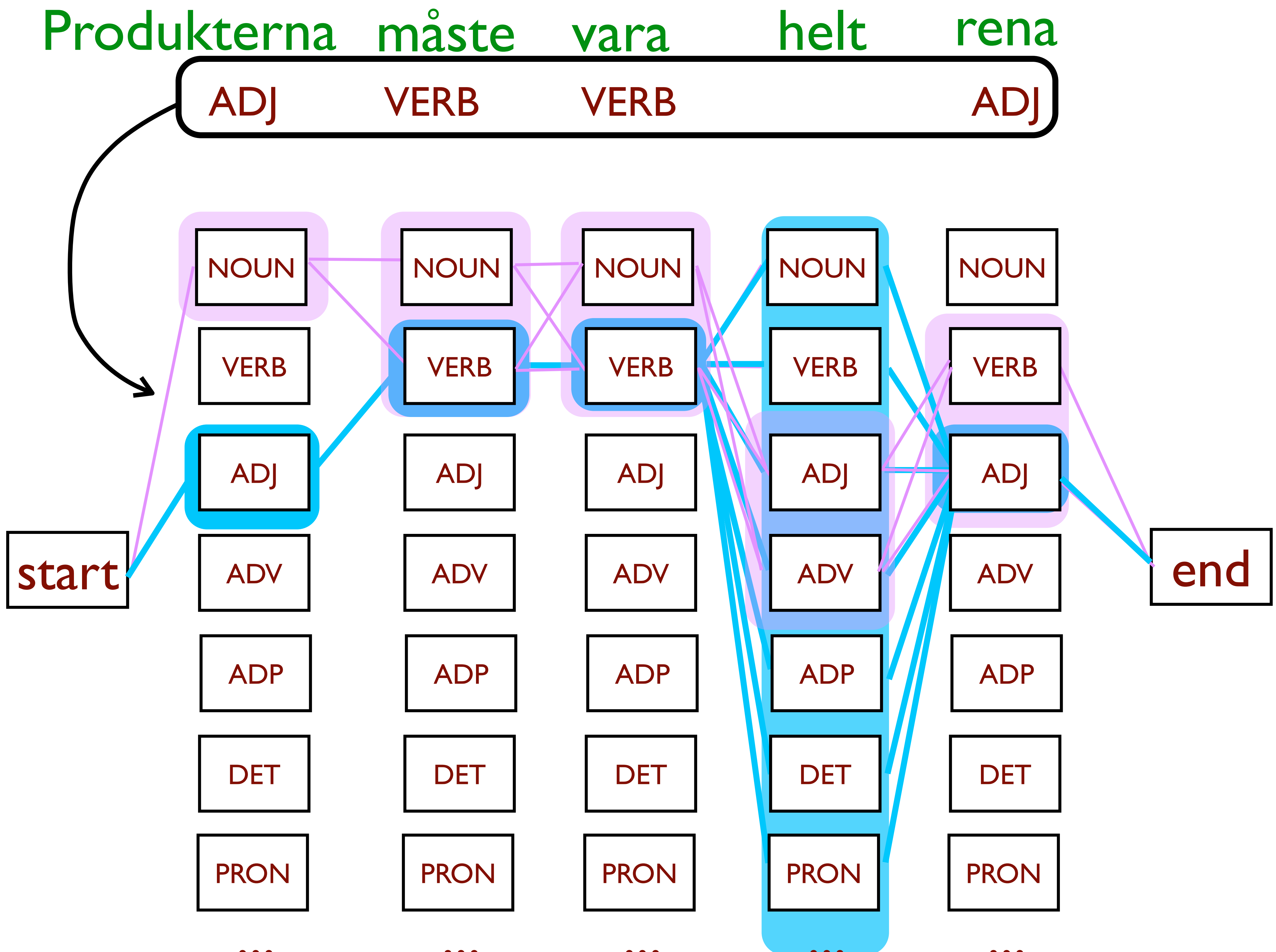
Coupled **Token** and **Type** Constraints



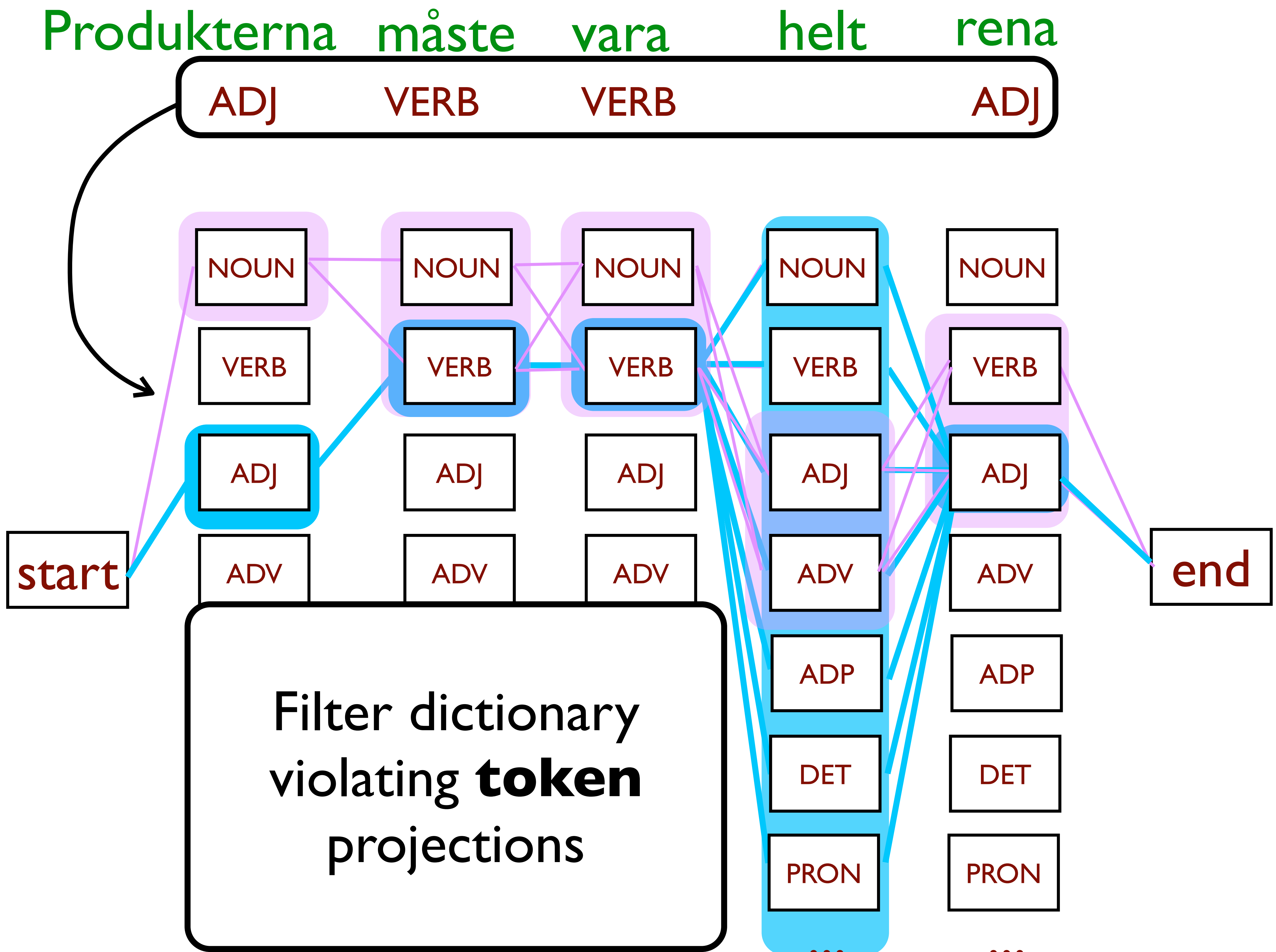
Coupled **Token** and **Type** Constraints



Coupled **Token** and **Type** Constraints

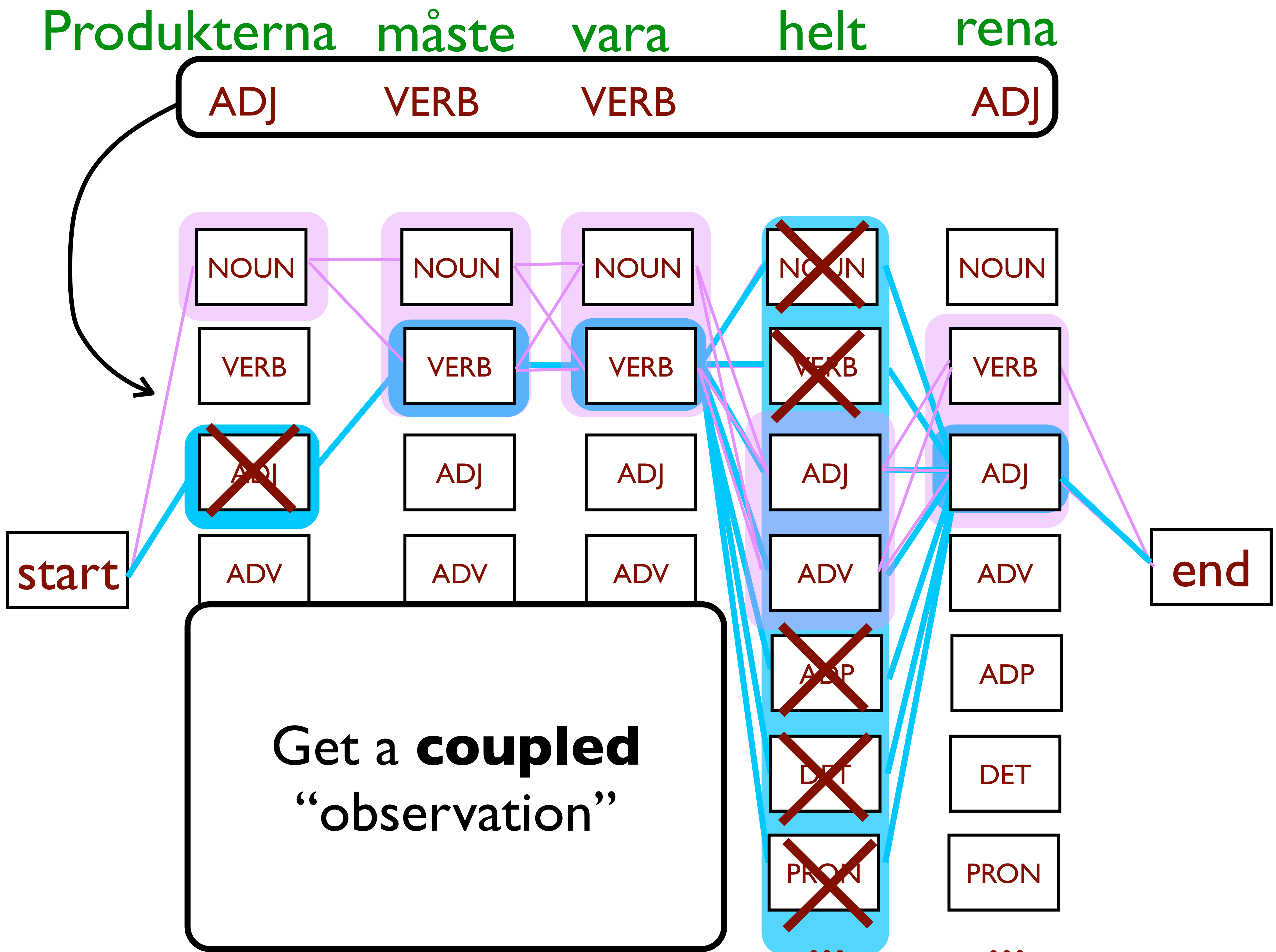


Coupled **Token** and **Type** Constraints

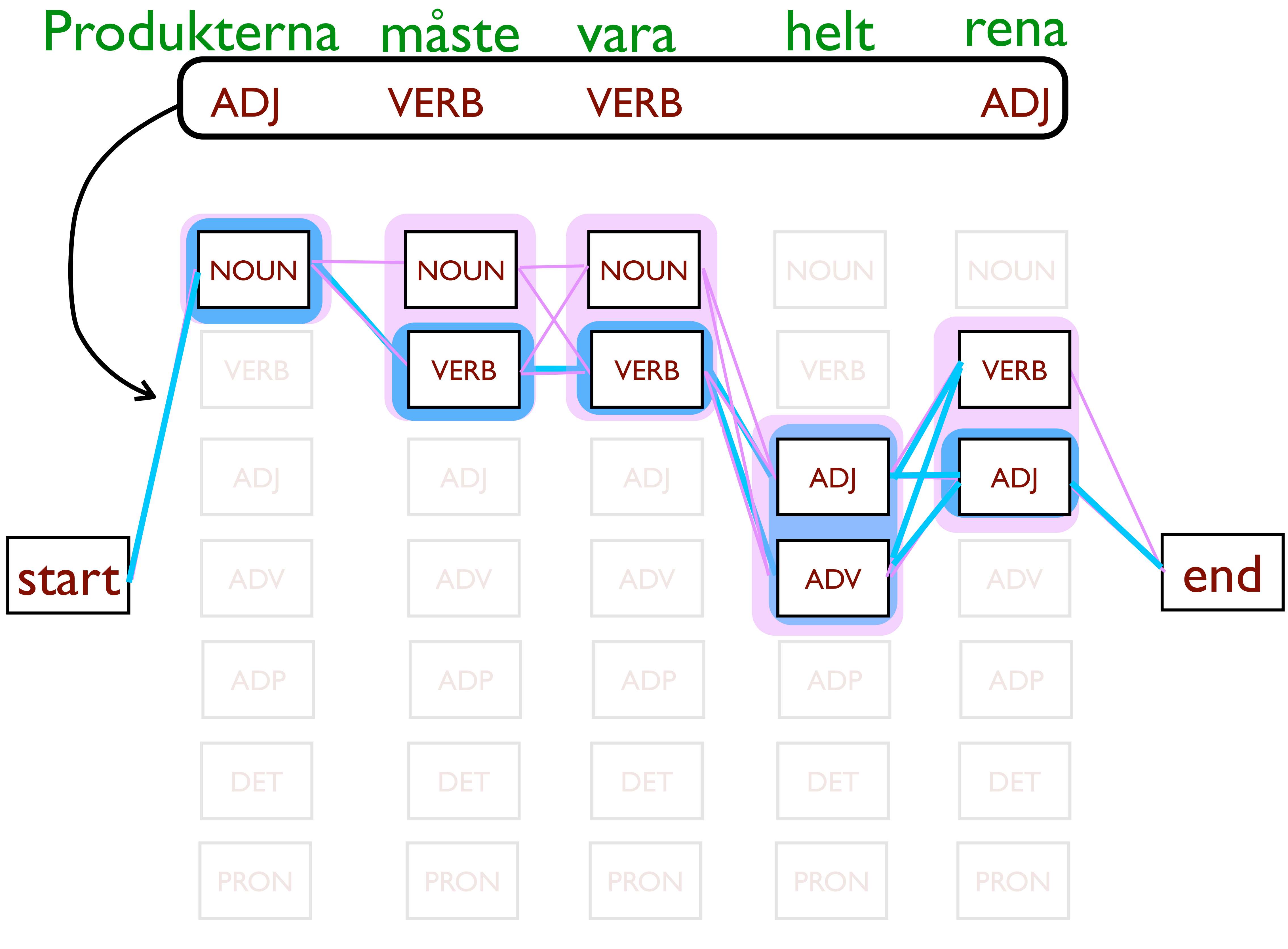


Filter dictionary violating **token** projections

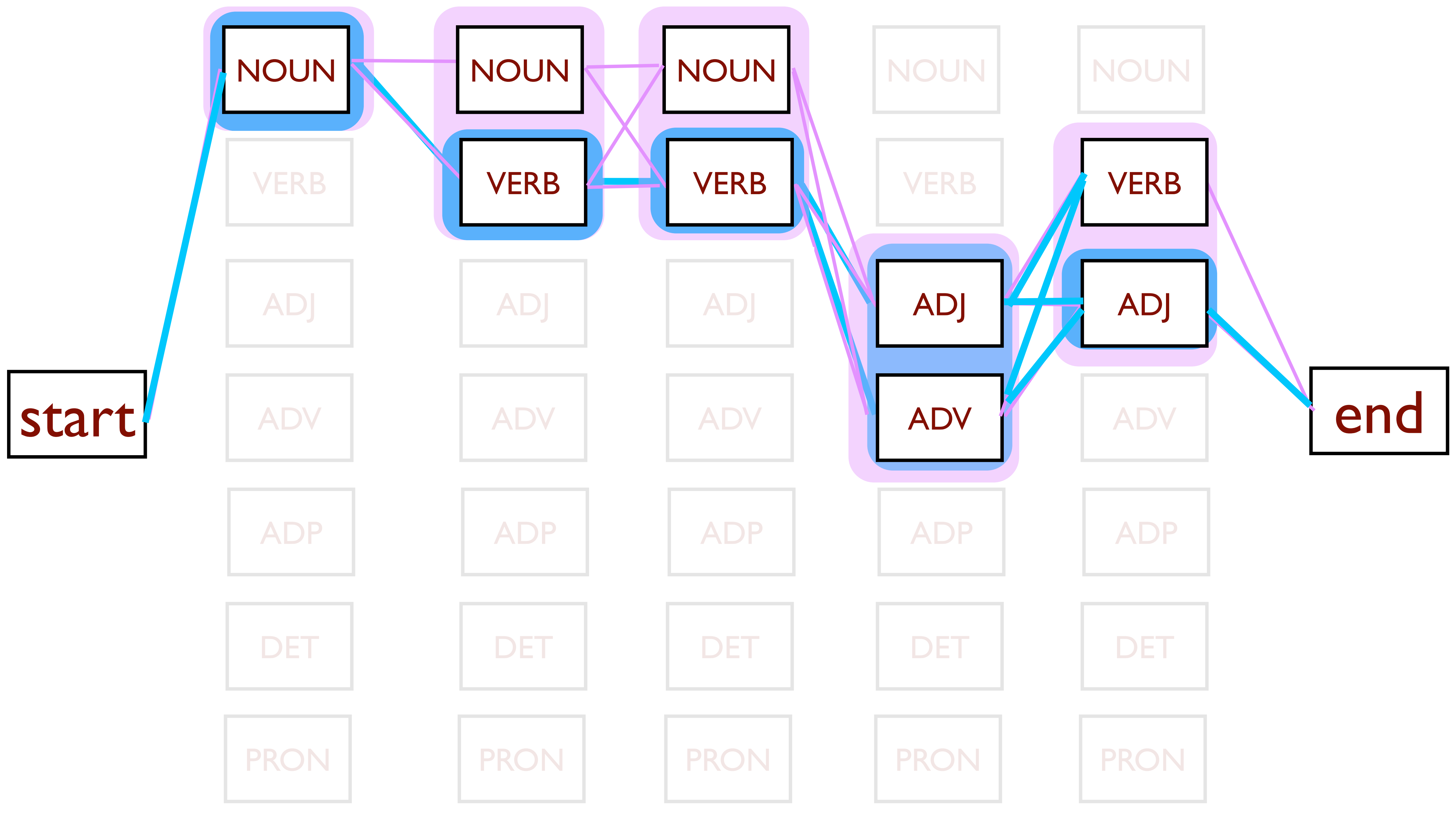
Coupled **Token** and **Type** Constraints



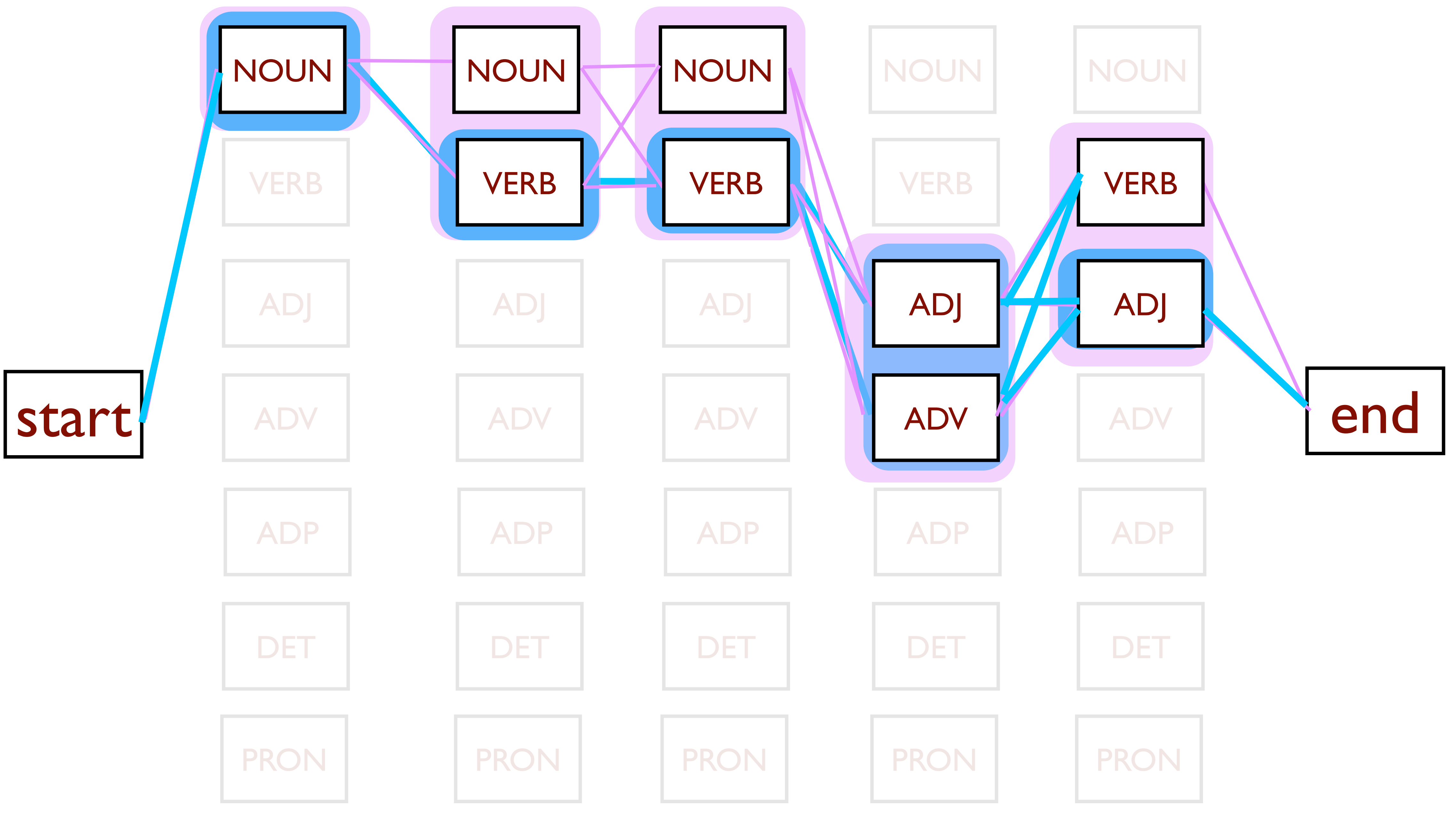
Coupled **Token** and **Type** Constraints



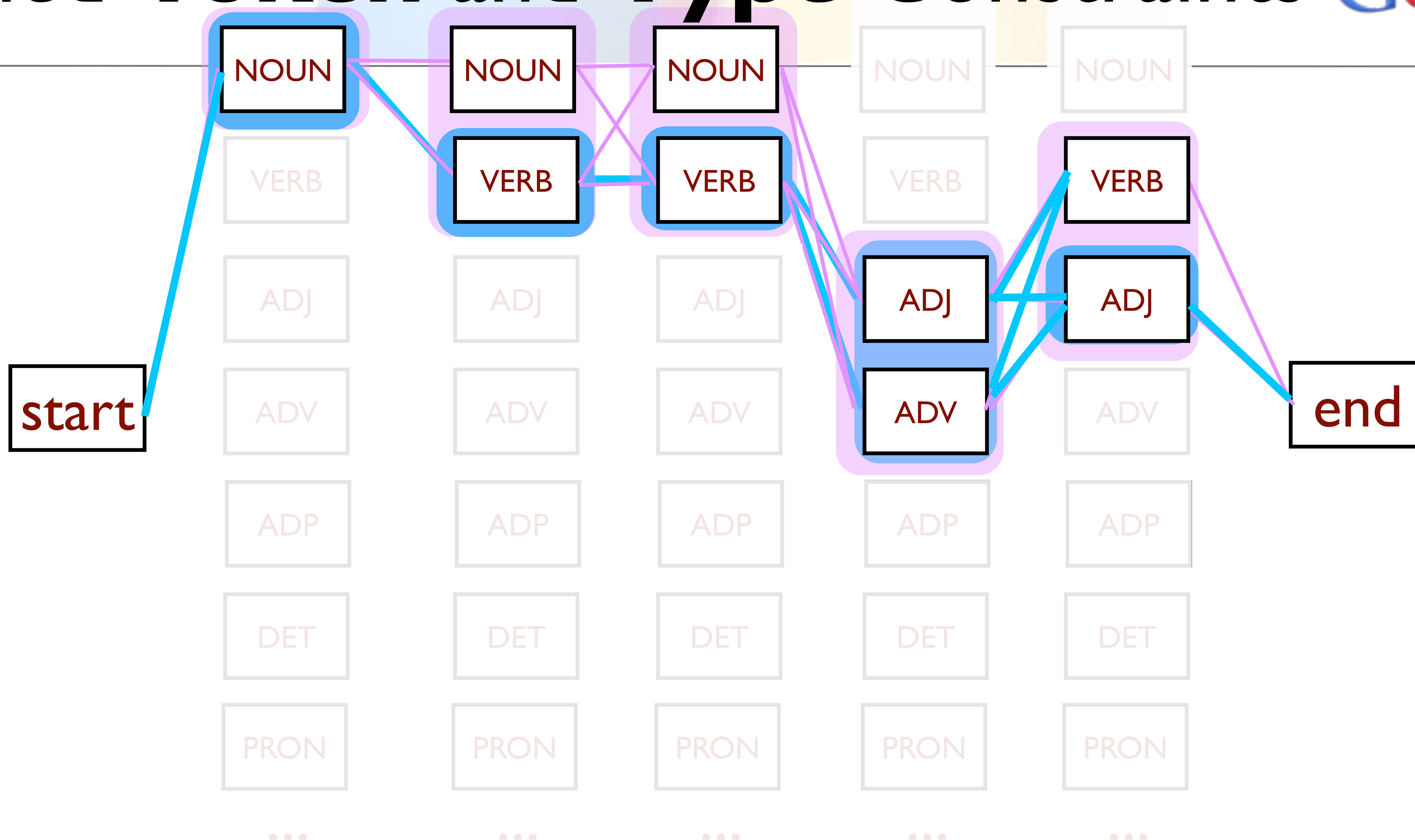
Coupled **Token** and **Type** Constraints



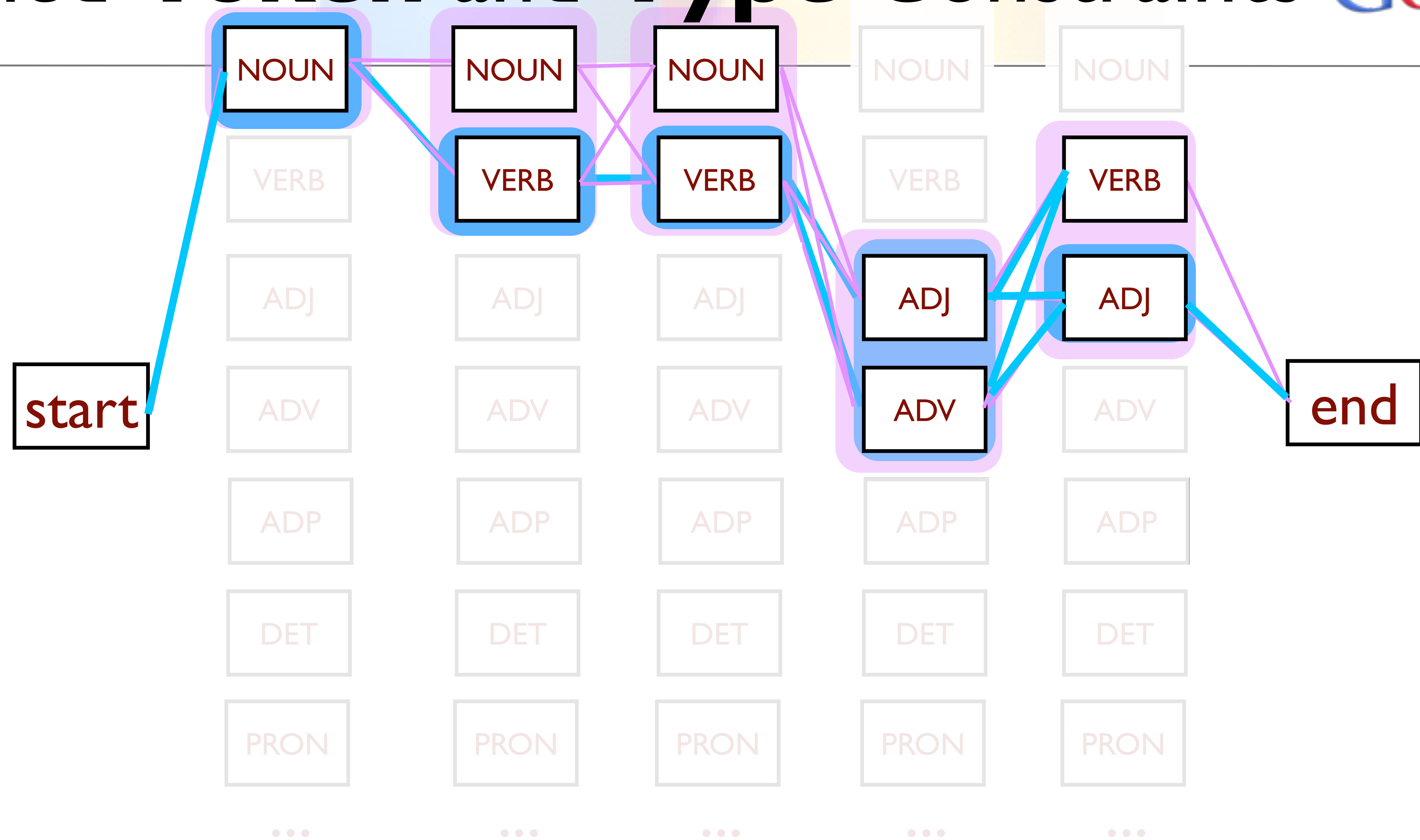
Coupled **Token** and **Type** Constraints



Coupled **Token** and **Type** Constraints

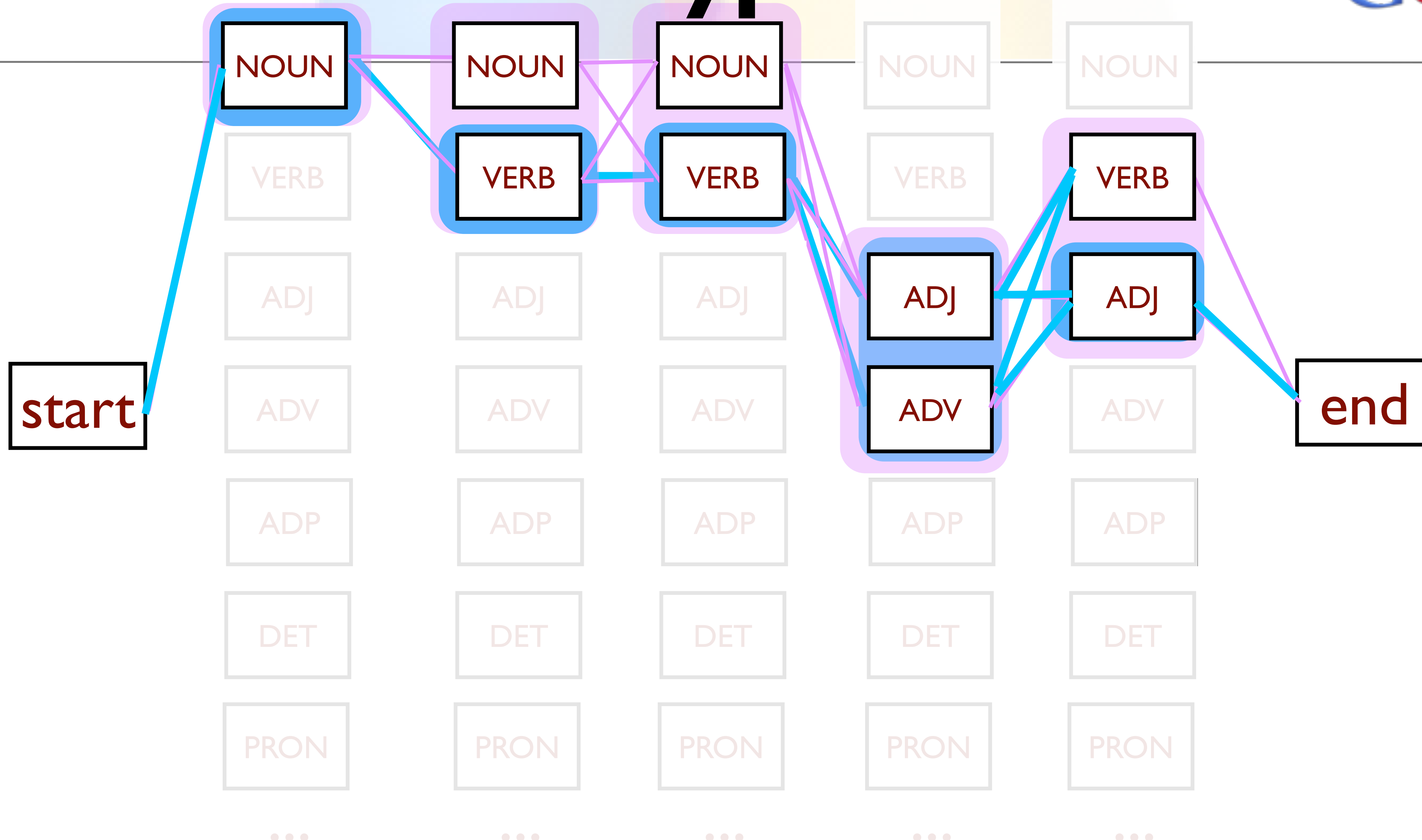


Coupled **Token** and **Type** Constraints



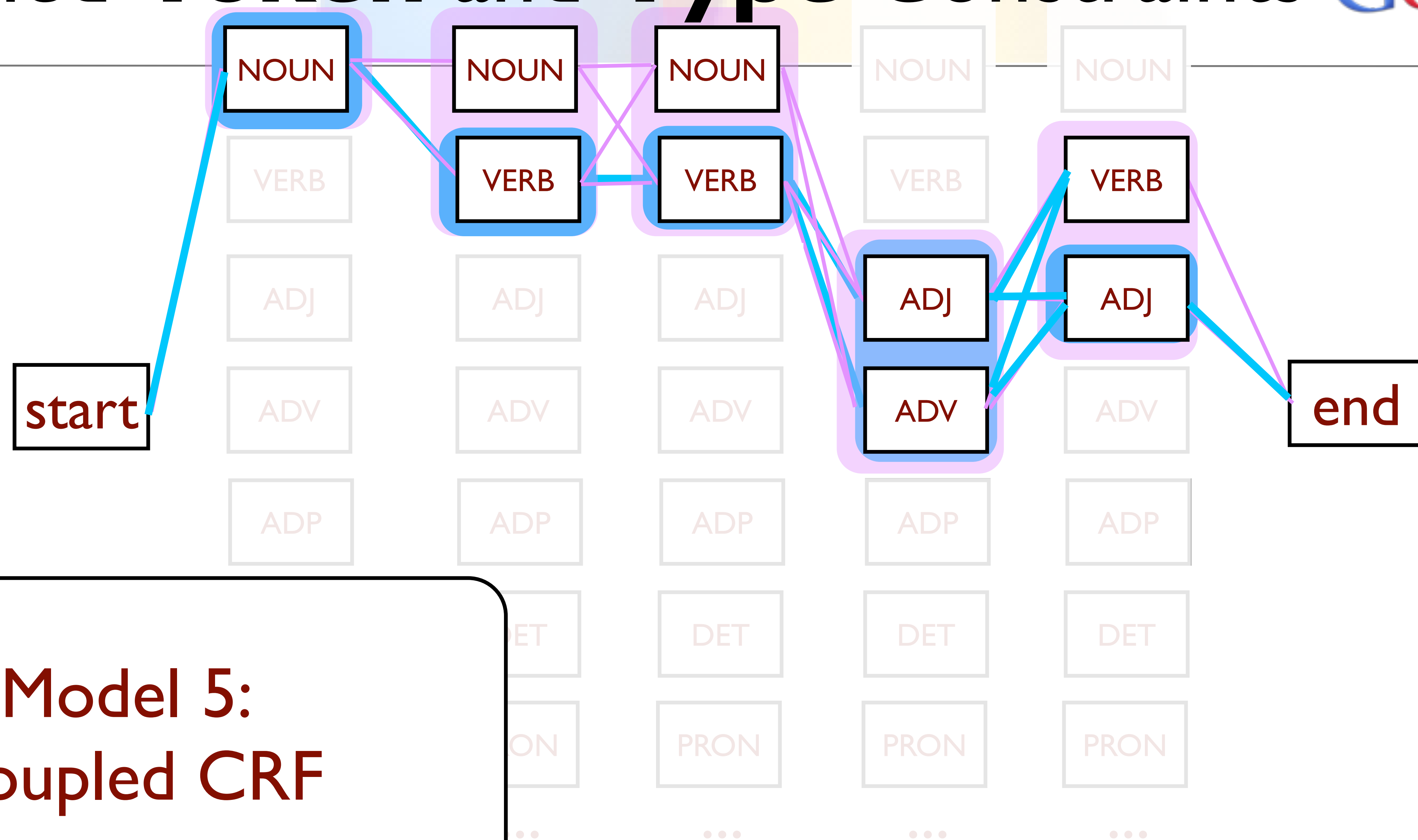
$$p(\boldsymbol{y} \in \mathrm{wiki}(\boldsymbol{x}) \frown \tilde{\boldsymbol{y}}|\boldsymbol{x}) = \frac{\displaystyle\sum_{\boldsymbol{y}\in\mathrm{wiki}(\boldsymbol{x})\frown\tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta}\cdot f(\boldsymbol{x},\boldsymbol{y})}{\displaystyle\sum_{\boldsymbol{y}'\in\mathrm{wiki}(\boldsymbol{x})} \exp \boldsymbol{\theta}\cdot f(\boldsymbol{x},\boldsymbol{y}')}$$

Coupled **Token** and **Type** Constraints



$$p(\boldsymbol{y} \in \mathrm{wiki}(\boldsymbol{x}) \frown \tilde{\boldsymbol{y}} | \boldsymbol{x}) = \frac{\sum_{\boldsymbol{y} \in \mathrm{wiki}(\boldsymbol{x}) \frown \tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathrm{wiki}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Coupled **Token** and **Type** Constraints

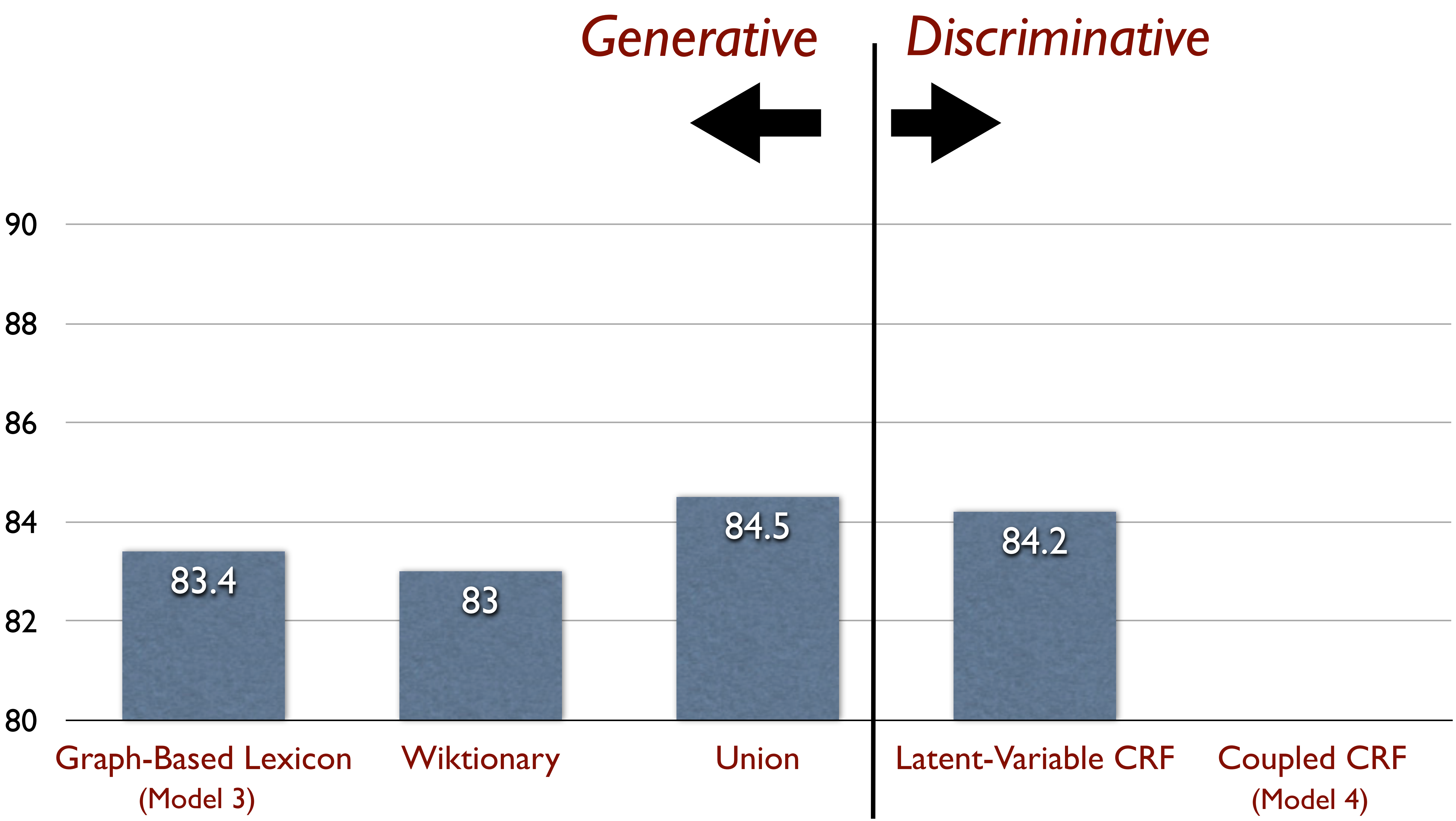


Model 5:
Coupled CRF

$$p(\boldsymbol{y} \in \text{wiki}(\boldsymbol{x}) \frown \tilde{\boldsymbol{y}} | \boldsymbol{x}) = \frac{\sum_{\boldsymbol{y} \in \text{wiki}(\boldsymbol{x}) \frown \tilde{\boldsymbol{y}}} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \text{wiki}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

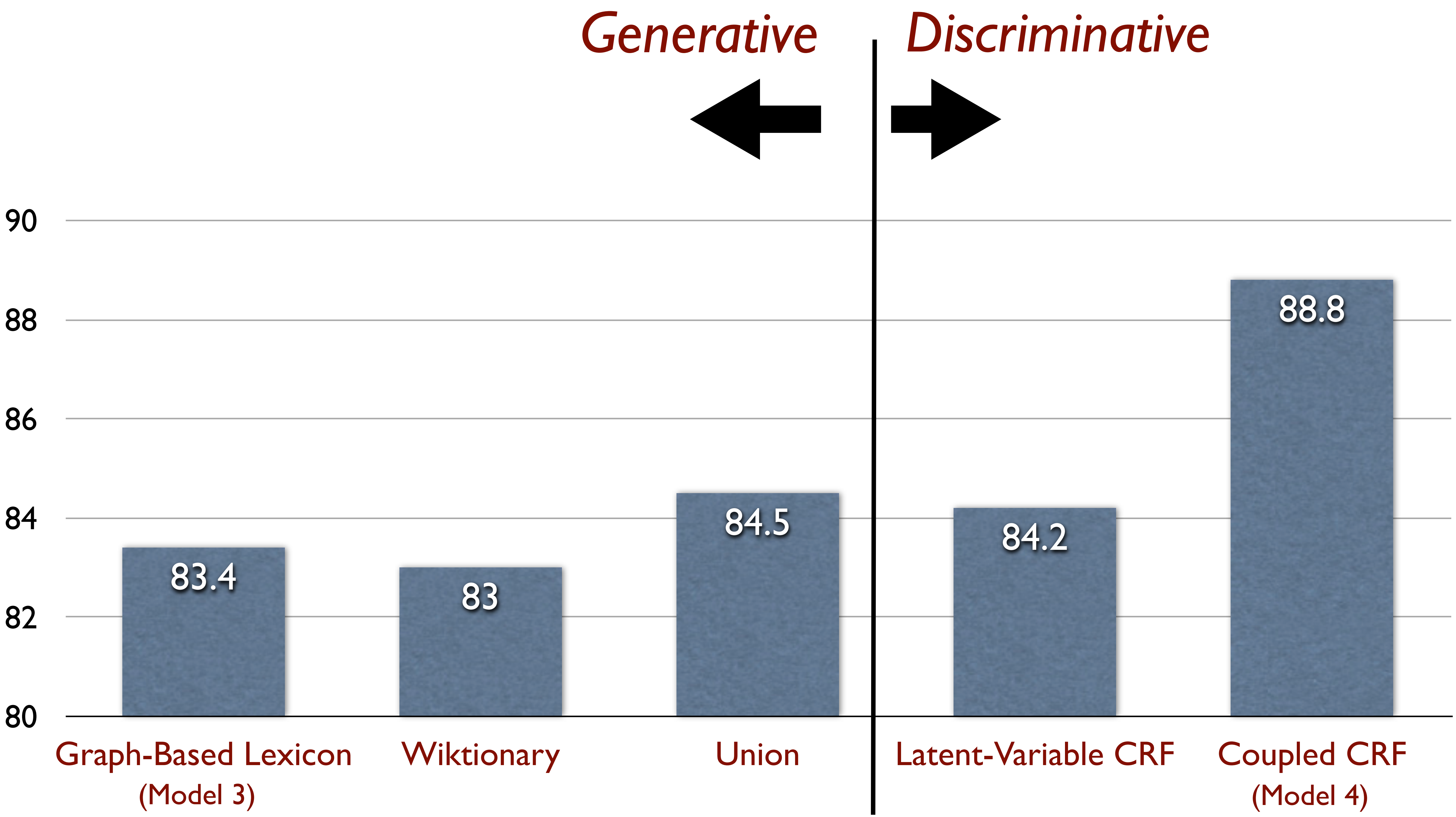
Model Comparison

Measured across 8 Indo-European Languages



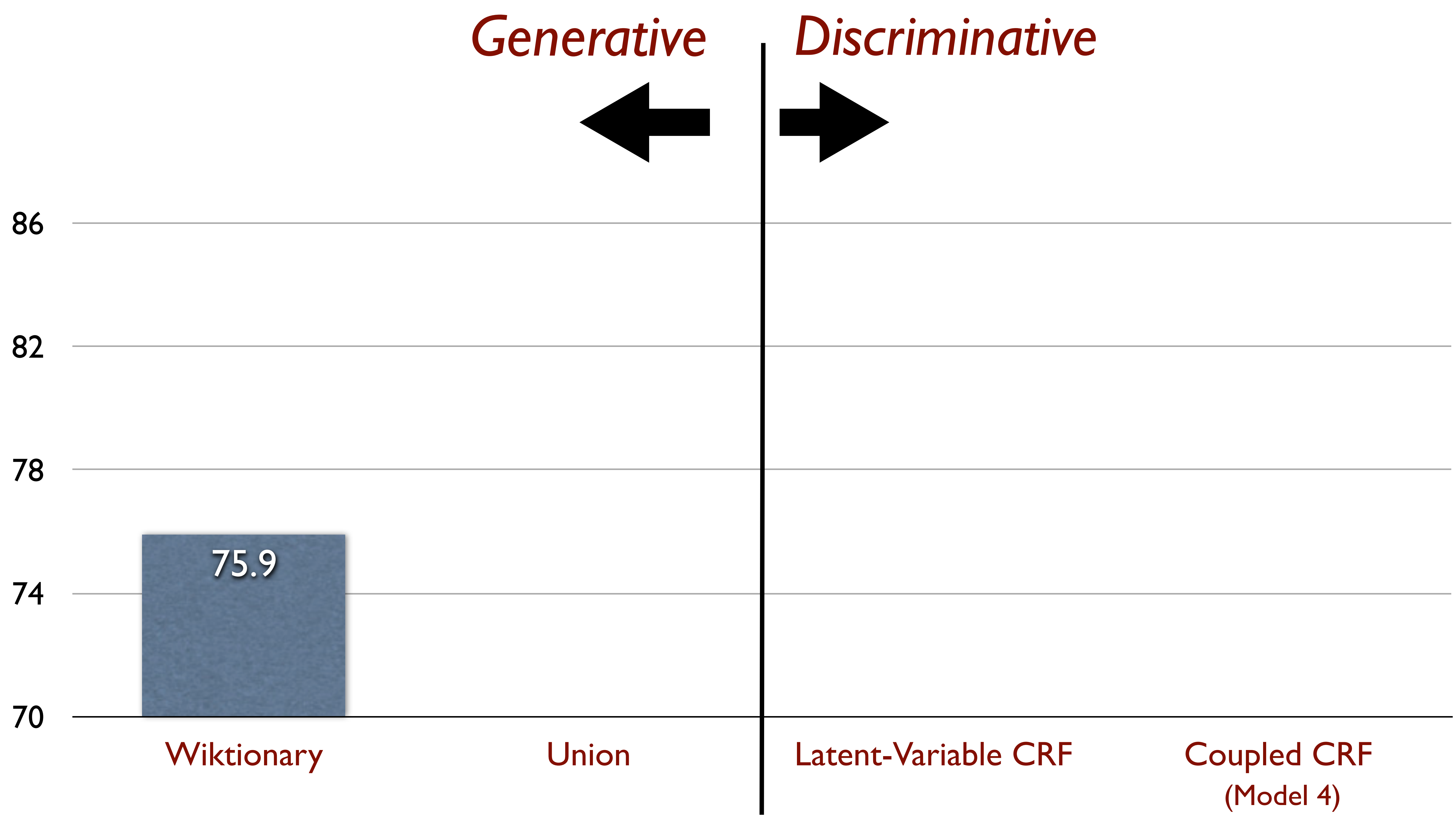
Model Comparison

Measured across 8 Indo-European Languages



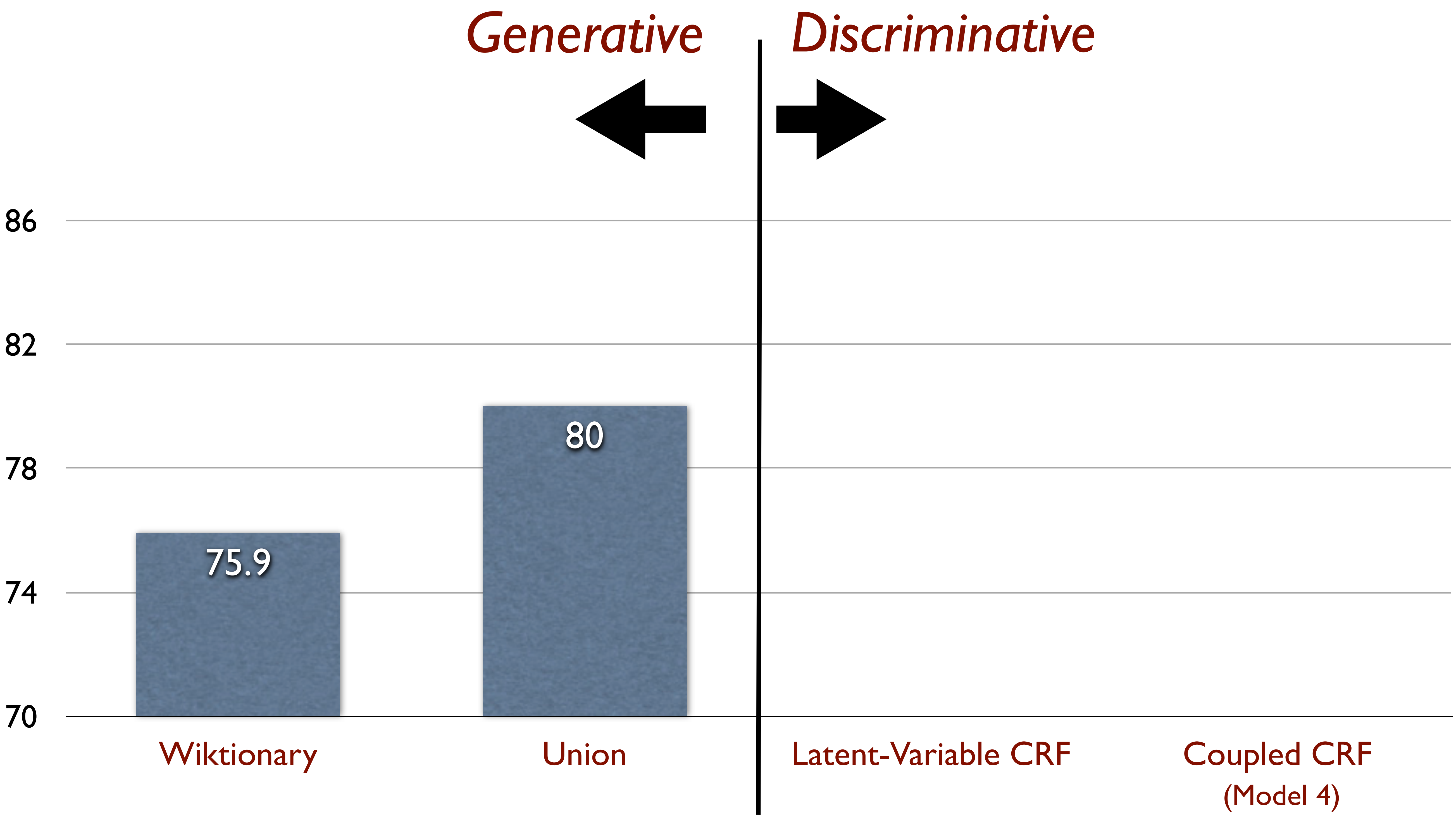
Model Comparison

Measured across 15 Languages



Model Comparison

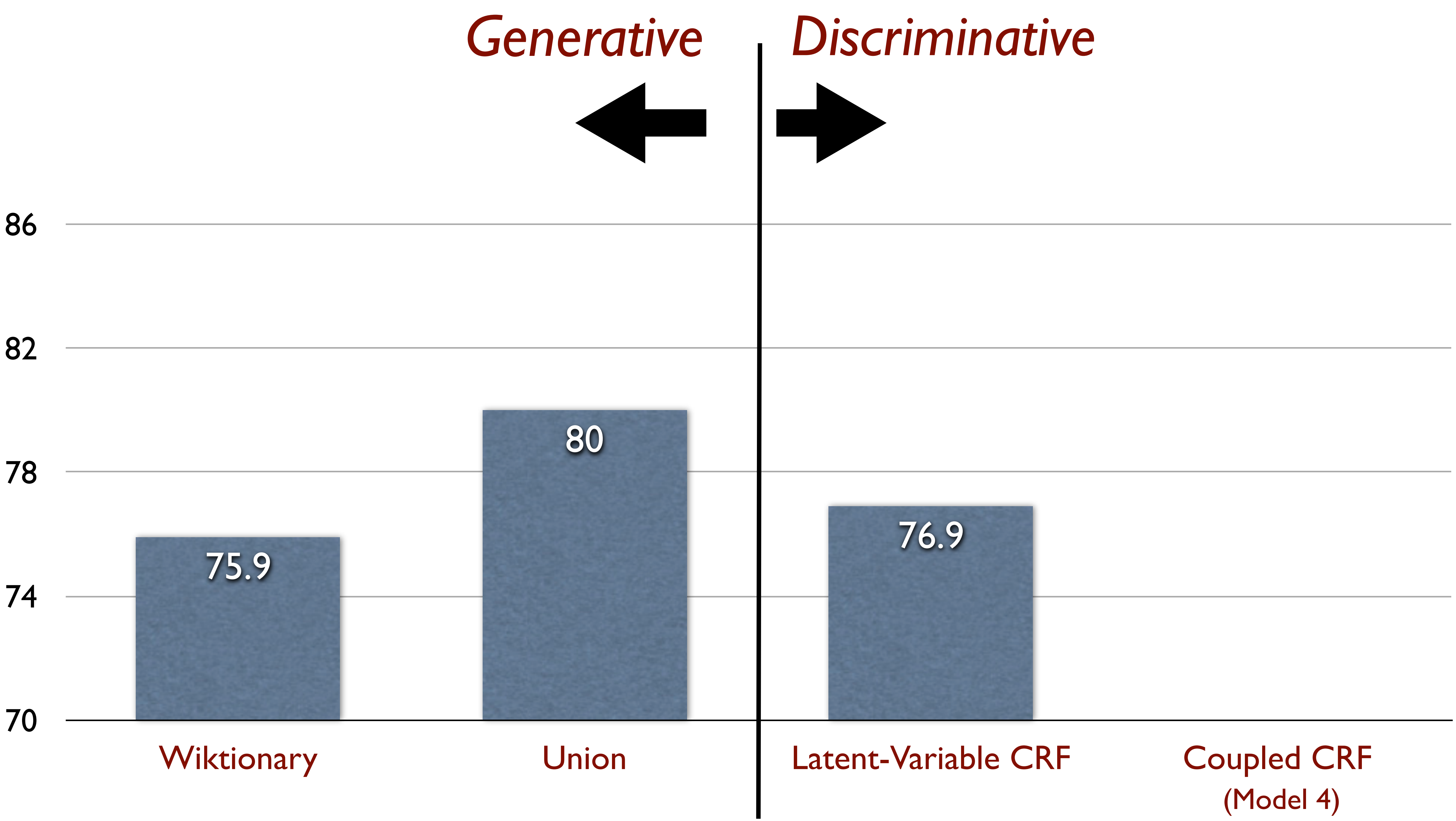
Measured across 15 Languages



Model Comparison

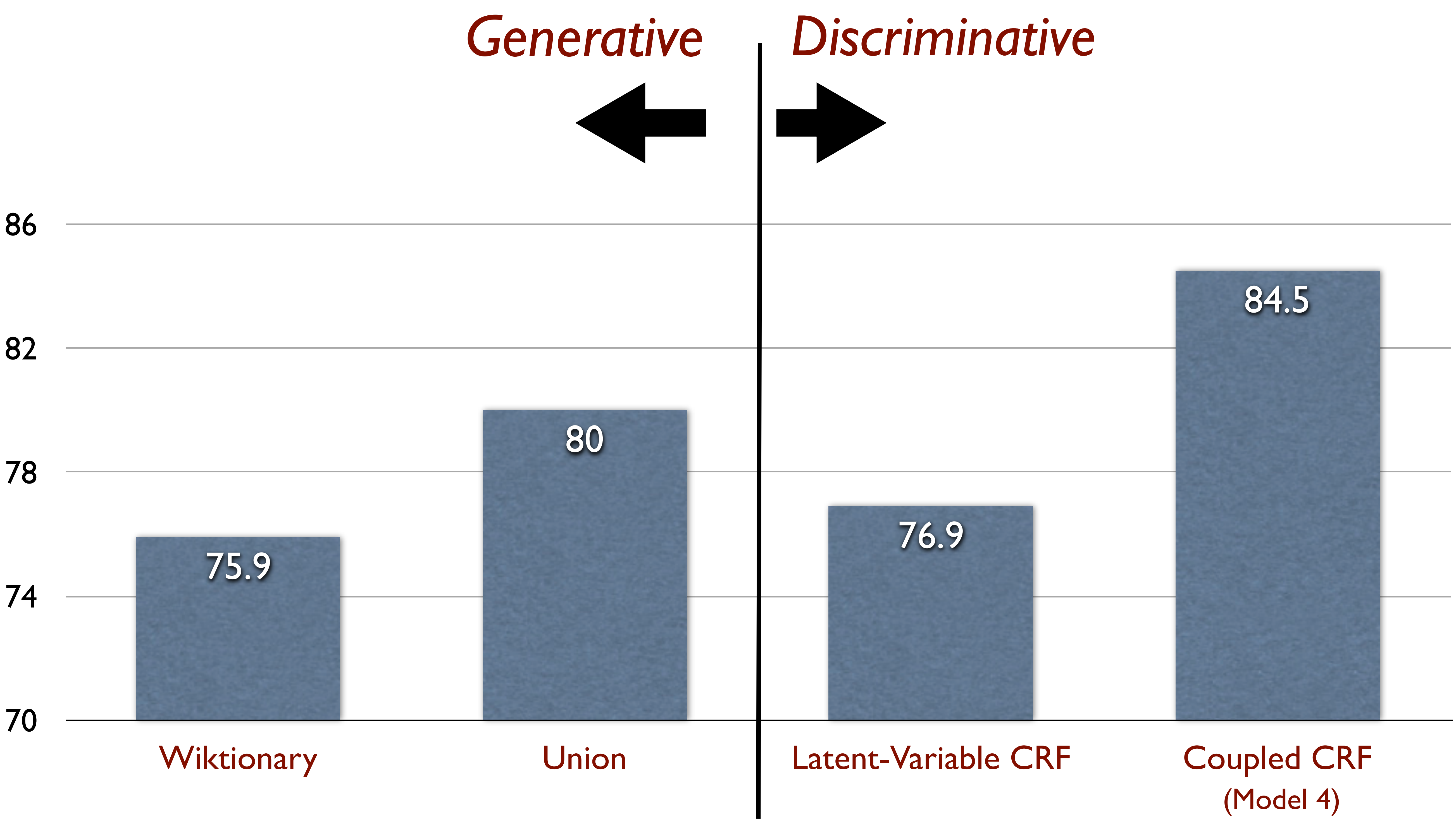
Google

Measured across 15 Languages



Model Comparison

Measured across 15 Languages



Talk Outline

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

Talk Outline

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

Posterior Regularization

In a discriminative model (CRF) :

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Posterior Regularization

In a discriminative model (CRF) :

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Entire dataset:

$$(\mathbf{X}, \mathbf{Y}) = ((\boldsymbol{x}_1, \boldsymbol{y}_1) \ldots (\boldsymbol{x}_n, \boldsymbol{y}_n))$$

Posterior Regularization

In a discriminative model (CRF) :

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{\theta} \cdot f(\boldsymbol{x}, \boldsymbol{y}')}$$

Entire dataset:

$$(\mathbf{X}, \mathbf{Y}) = ((\boldsymbol{x}_1, \boldsymbol{y}_1) \ldots (\boldsymbol{x}_n, \boldsymbol{y}_n))$$

Maximum conditional likelihood:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma \|\boldsymbol{\theta}\|$$

Posterior Regularization

Supervised Objective:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma\|\boldsymbol{\theta}\|$$

Posterior Regularization

Supervised Objective:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma||\boldsymbol{\theta}||$$

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma||\boldsymbol{\theta}|| + \text{constant}$$

Posterior Regularization

Supervised Objective:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma\|\boldsymbol{\theta}\|$$

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma\|\boldsymbol{\theta}\| + \text{constant}$$

Distribution with probability of labeled data = 1

Posterior Regularization

Supervised Objective:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma||\boldsymbol{\theta}||$$

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma||\boldsymbol{\theta}|| + \text{constant}$$

Distribution with probability of labeled data = 1

What if we only have partial labeling?

Posterior Regularization

Google

Supervised Objective:

Distribution with probability of labeled data = 1

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma\|\boldsymbol{\theta}\| + \text{constant}$$

Posterior Regularization

Supervised Objective:

Distribution with probability of labeled data = 1

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma\|\boldsymbol{\theta}\| + \text{constant}$$

Objective with Partial Labelings:

Rather than a single q, we have entire set $Q = \{q\}$

Posterior Regularization

Supervised Objective:

Distribution with probability of labeled data = 1

$$= \min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma\|\boldsymbol{\theta}\| + \text{constant}$$

Objective with Partial Labelings:

Rather than a single q, we have entire set $Q = \{q\}$

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$
$$= \max_{\boldsymbol{\theta}} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) - \gamma\|\boldsymbol{\theta}\| - \text{constant}$$

Posterior Regularization

Example:

The tag "VERB" must appear at least 10% of the time.

$$\phi(\mathbf{X}, \mathbf{Y}) = \sum_{y_i \in \mathbf{Y}} \begin{cases} \frac{1}{|\mathbf{Y}|} & \text{if } y_i = \text{VERB} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{Q} = \{q(\mathbf{Y}) : \mathbf{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \geq 0.1\}$$

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

$$= \max_{\boldsymbol{\theta}} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) - \gamma\|\boldsymbol{\theta}\| - \text{constant}$$

Posterior Regularization

Example:

The tag "VERB" must appear at least 10% of the time.

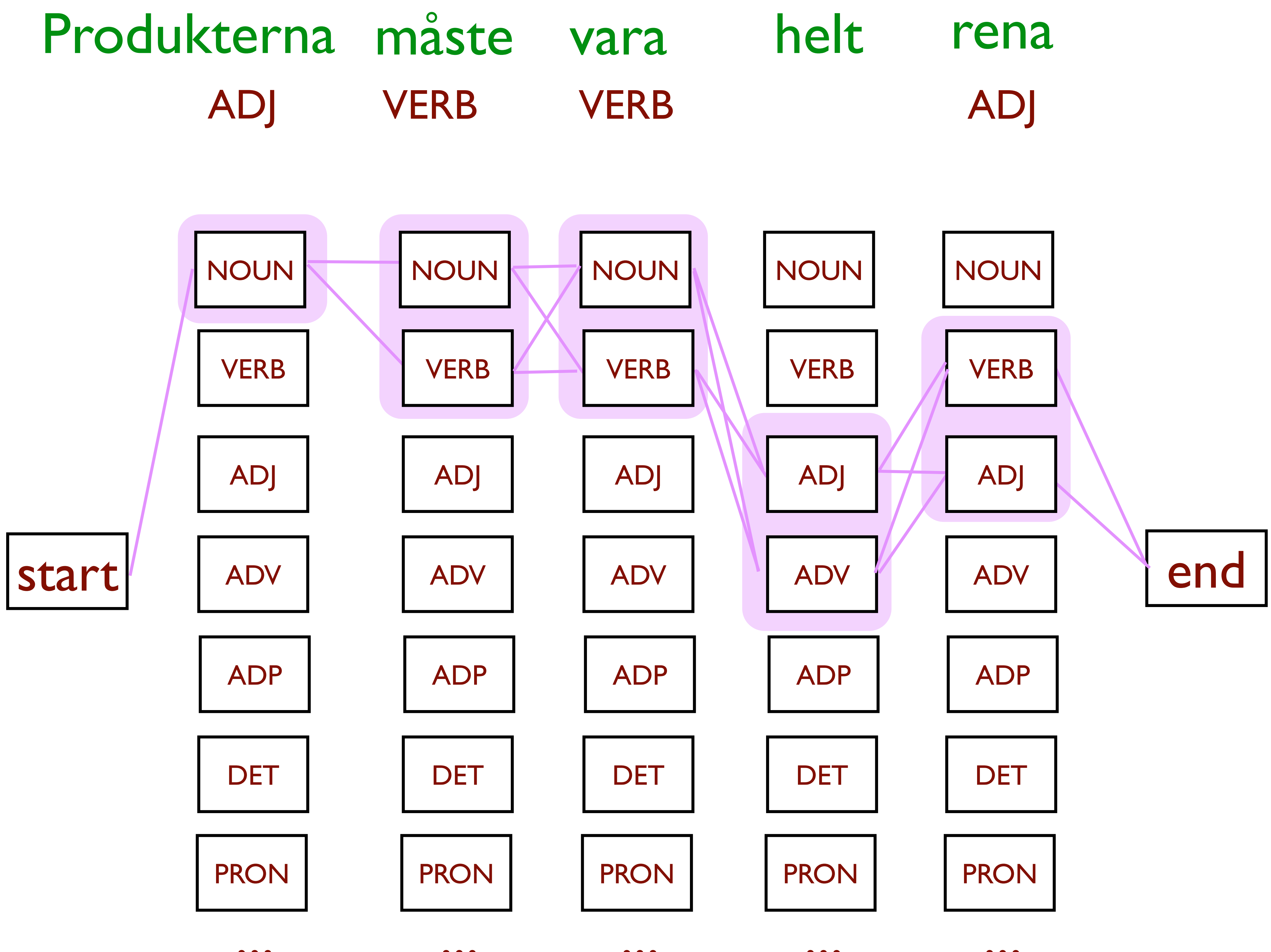
$$\boxed{\phi(\mathbf{X}, \mathbf{Y}) = \sum_{y_i \in \mathbf{Y}} \begin{cases} \frac{1}{|\mathbf{Y}|} & \text{if } y_i = \text{VERB} \\ 0 & \text{otherwise} \end{cases}}$$

a constraint feature

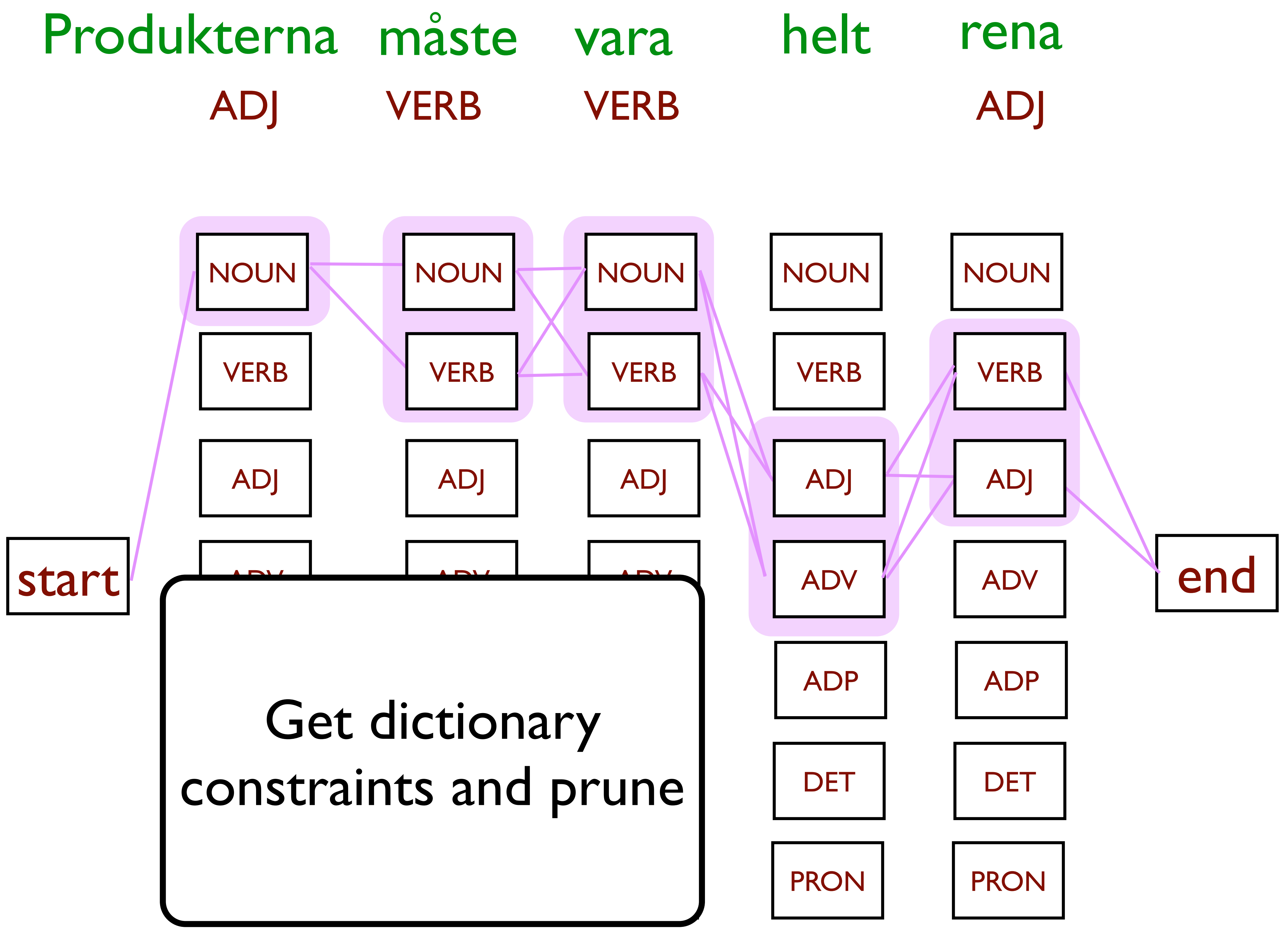
$$\mathcal{Q} = \{q(\mathbf{Y}) : \mathbf{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \geq 0.1\}$$

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$
$$= \max_{\boldsymbol{\theta}} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q|p(\mathbf{Y}|\mathbf{X})) - \gamma\|\boldsymbol{\theta}\| - \text{constant}$$

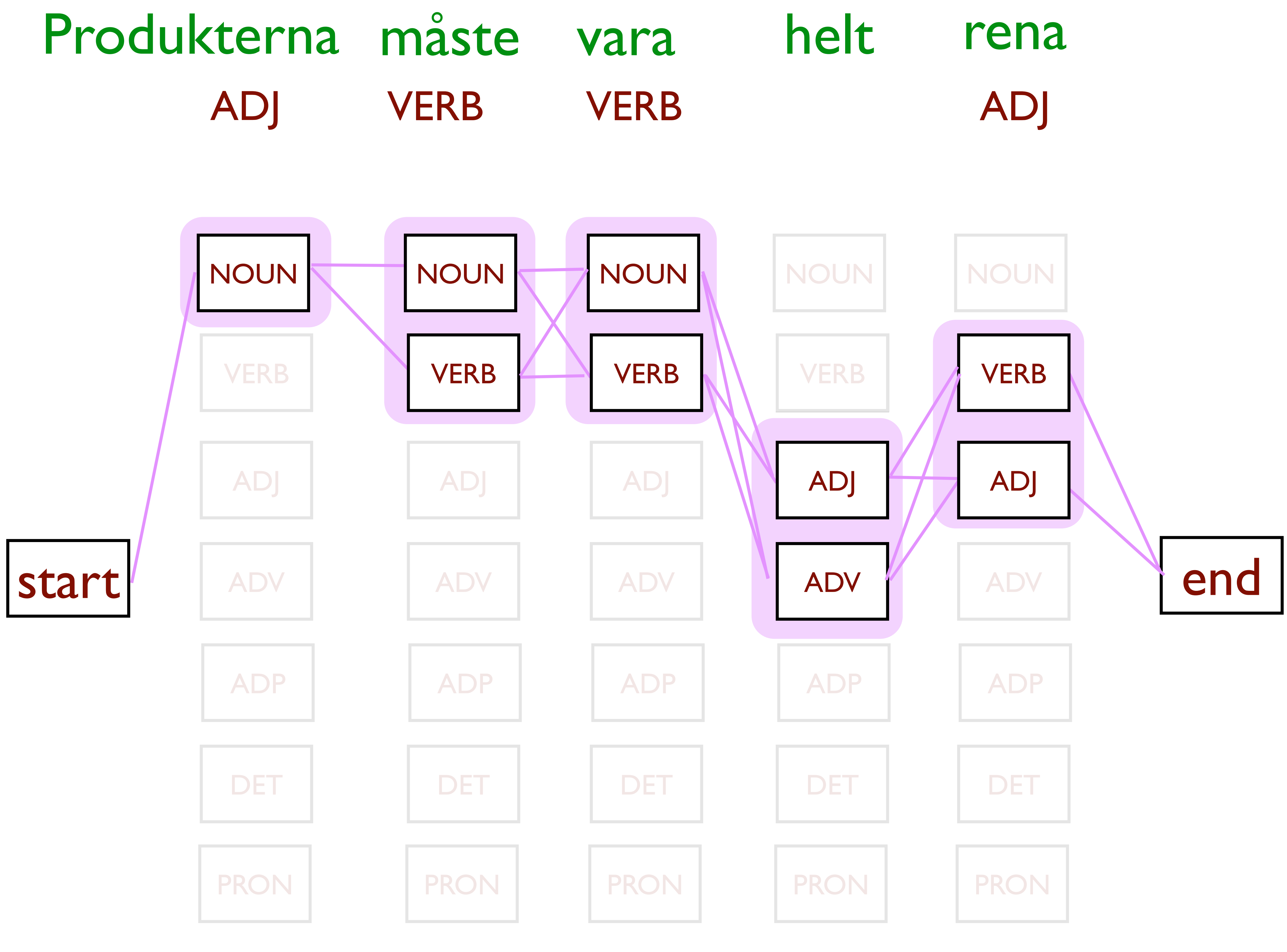
Posterior Regularization and Projection



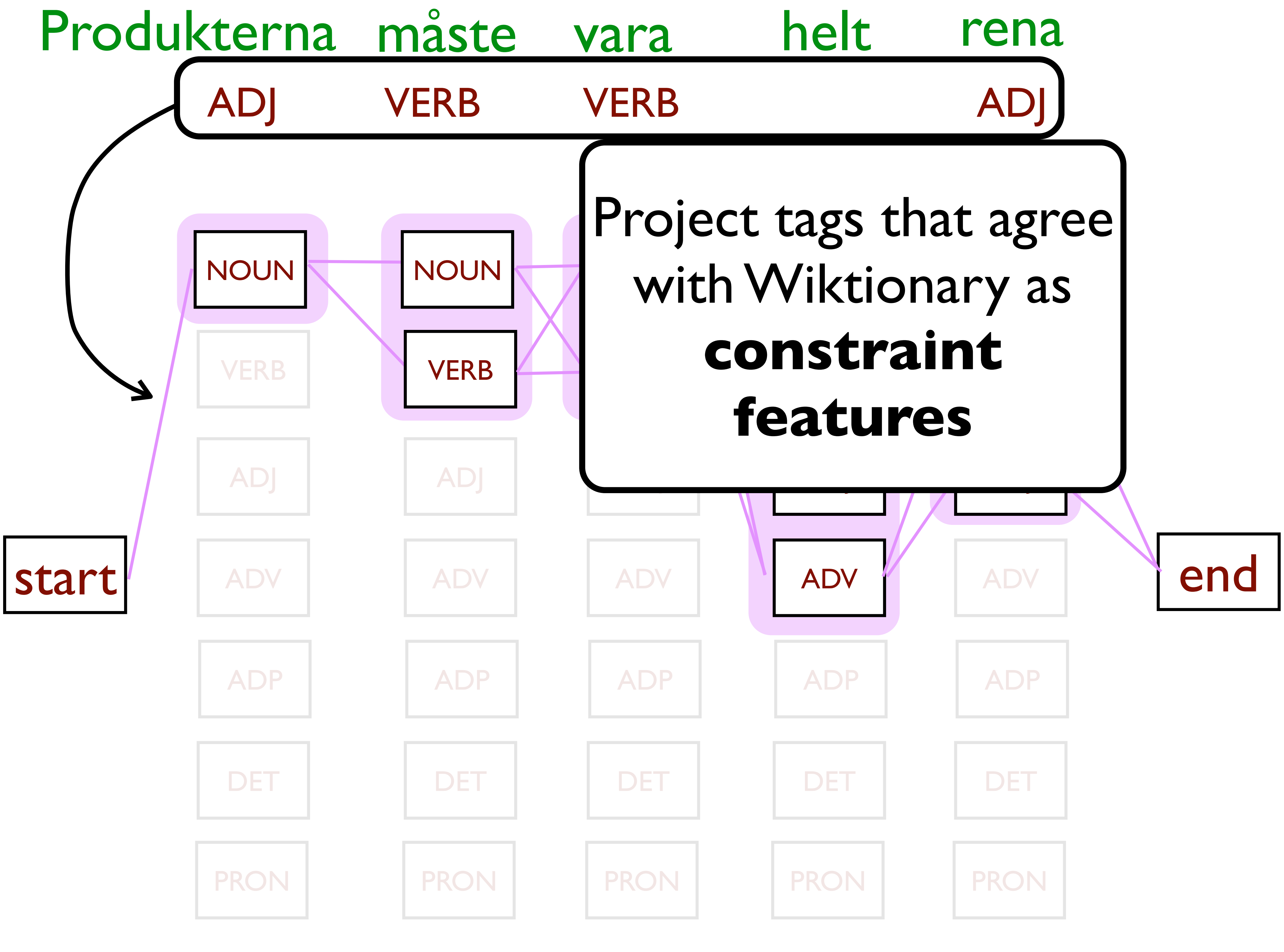
Posterior Regularization and Projection



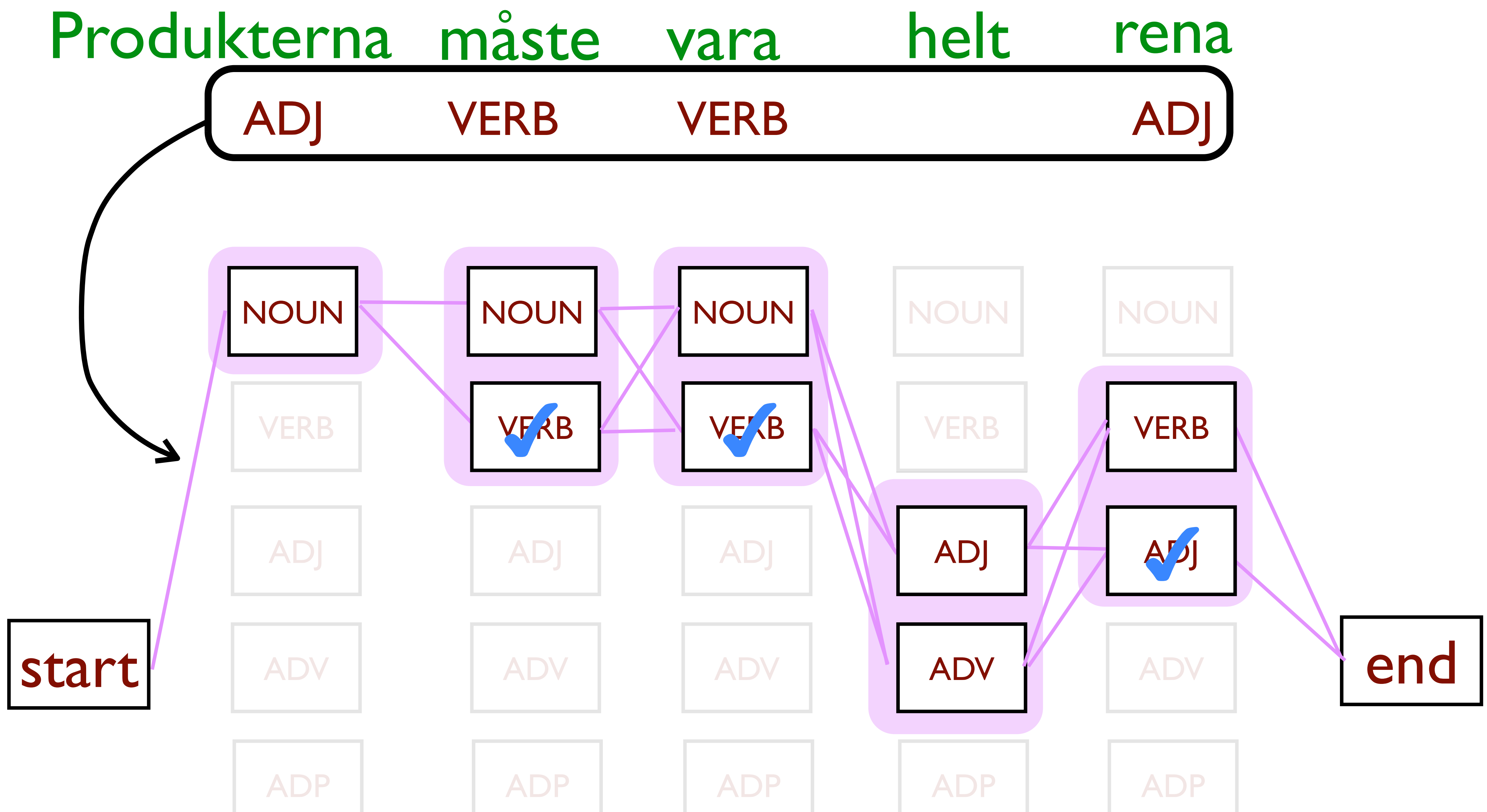
Posterior Regularization and Projection



Posterior Regularization and Projection



Posterior Regularization and Projection



$$\mathcal{Q} = \{q(\mathbf{Y}) : \mathbf{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \geq \mathbf{b}\}$$

$$\phi(\mathbf{X}, \mathbf{Y}) = \text{number of } \checkmark \in \mathbf{Y}$$

Learning

Objective:

$$\max_{\theta} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q | p_\theta) - \gamma \|\theta\|$$

Learning

Objective:

$$\max_{\theta} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q|p_\theta) - \gamma\|\theta\|$$

Dual with respect to q :

$$\max_{\theta} \min_{\lambda} \mathbf{b} \cdot \lambda + \sum_{\mathbf{Y}} p_\theta(\mathbf{Y}|\mathbf{X}) e^{-\lambda \cdot \phi(\mathbf{X},\mathbf{Y})} - \gamma\|\theta\|$$

Learning

Objective:

$$\max_{\theta} \max_{q \in \mathcal{Q}} -\mathcal{D}_{\mathrm{KL}}(q | p_{\theta}) - \gamma \|\theta\|$$

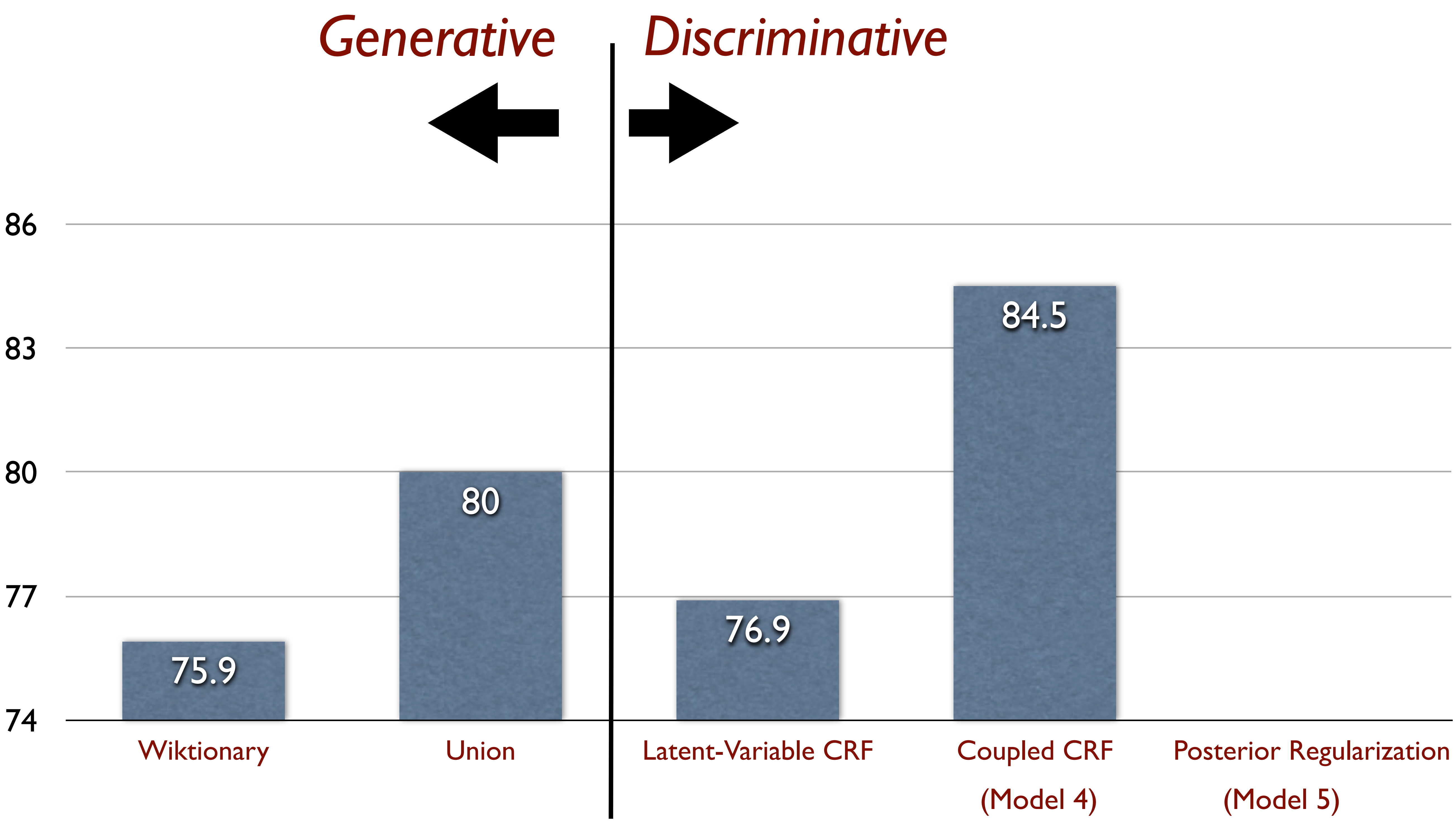
Dual with respect to q :

$$\max_{\theta} \min_{\lambda} \mathbf{b} \cdot \lambda + \sum_{\mathbf{Y}} p_{\theta}(\mathbf{Y}|\mathbf{X}) e^{-\lambda \cdot \phi(\mathbf{X}, \mathbf{Y})} - \gamma \|\theta\|$$

Optimized using alternating stochastic gradient descent

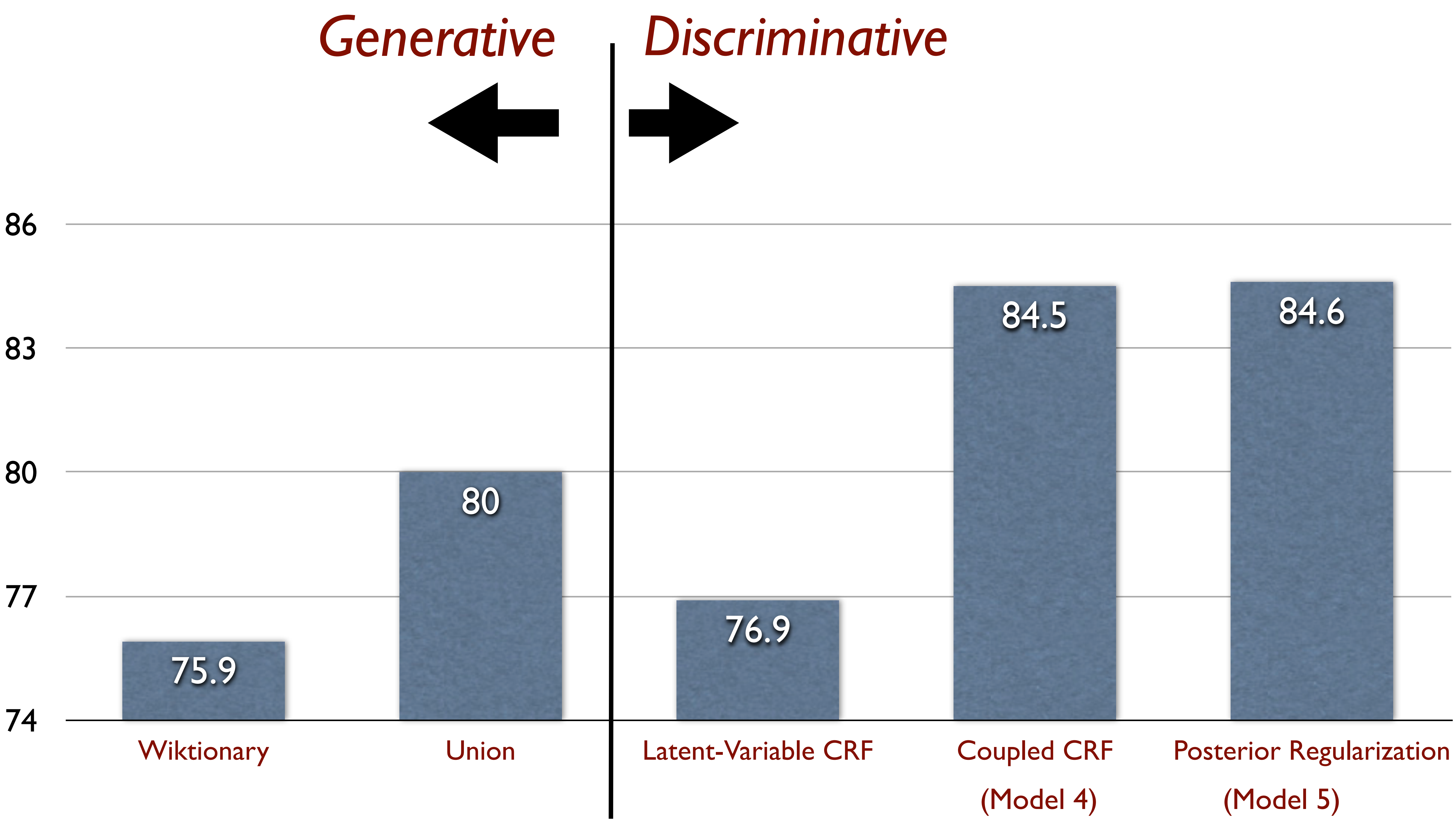
Model Comparison

Measured across 15 Languages



Model Comparison

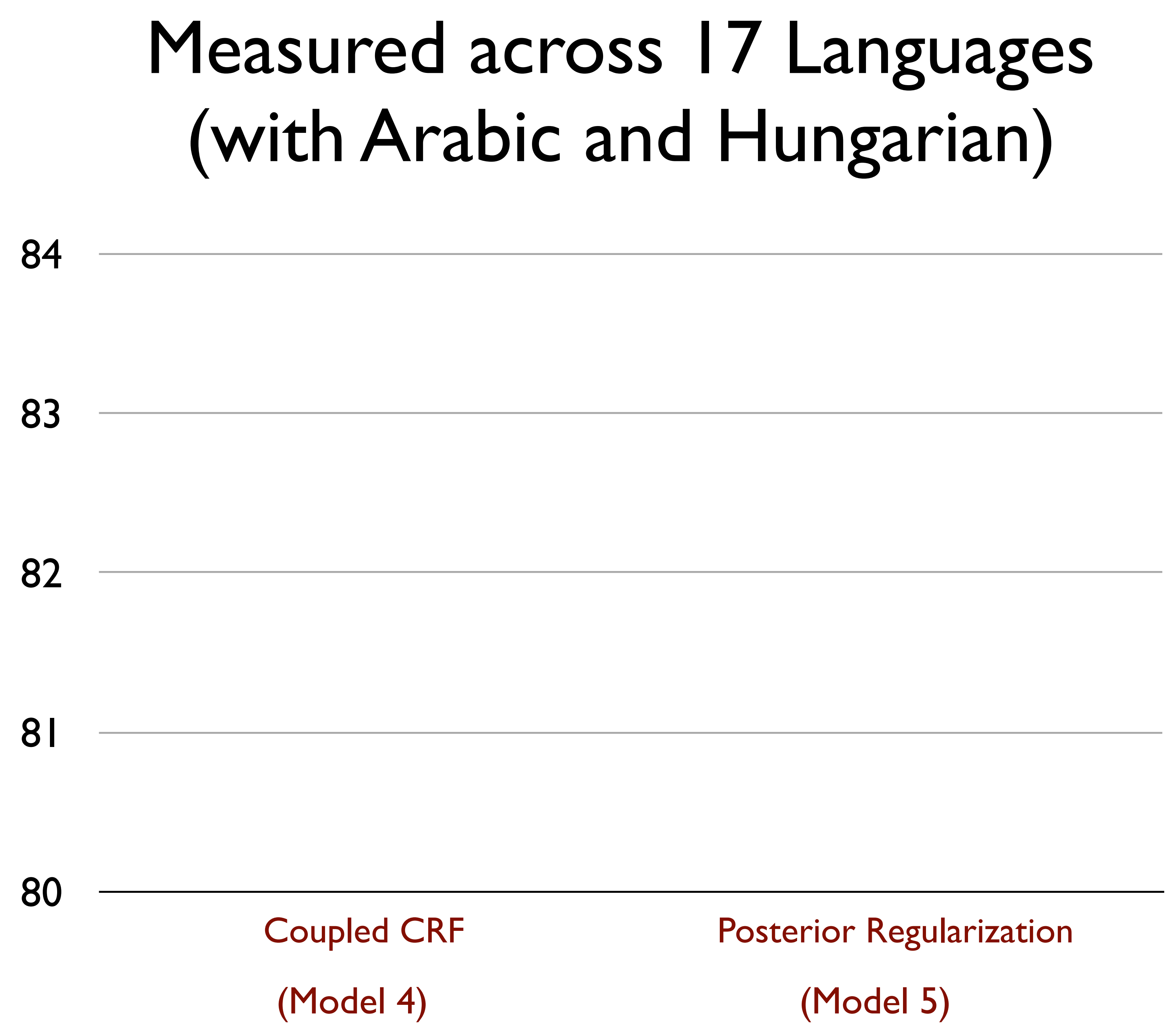
Measured across 15 Languages



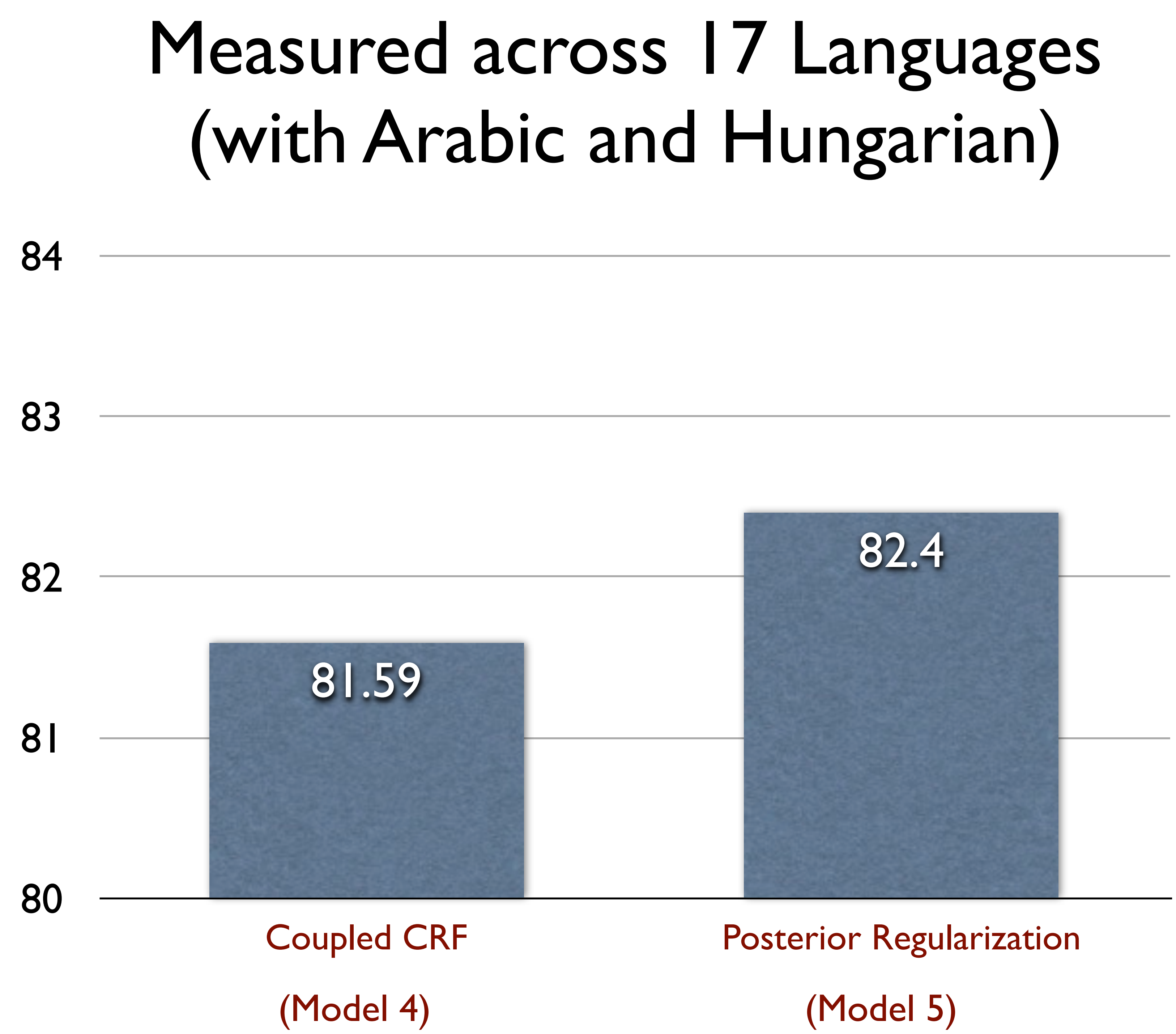
Model Comparison

Measured across 17 Languages
(with Arabic and Hungarian)

Model Comparison



Model Comparison



Talk Outline

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences and Dictionaries

Generalizing above with *Posterior Regularization*

See paper for experiments on named-entity segmentation

Talk Outline

- Cross-lingual learning for part-of-speech tagging
  - Use of translated data from English to target languages
  - Crowdsourced dictionaries
- A word about cross-lingual learning for syntactic parsing

Cross-Lingual Learning of Parsers

Cross-Lingual Learning of Parsers

- Large body of work
  - Learning of parsers jointly using parallel data
    - Snyder, Naseem and Barzilay (2009)
  - Parameter transfer from source languages to a target language
    - Cohen, Das, Smith (2011)
    - Naseem, Barzilay and Globerson (2012)
  - Target language adaptation of projected parsers
    - Täckström, McDonald and Nivre (2013)

Cross-Lingual Learning of Parsers

- Large body of work
  - Learning of parsers jointly using parallel data
    - Snyder, Naseem and Barzilay (2009)
  - Parameter transfer from source languages to a target language
    - Cohen, Das, Smith (2011)
    - Naseem, Barzilay and Globerson (2012)
  - Target language adaptation of projected parsers
    - Täckström, McDonald and Nivre (2013)

Constrained unsupervised learning

Summary

- Several methods for learning part-of-speech tagger systems without any direct supervision.
- Using large amounts of parallel data that are available for tens of languages.
- Using Wiktionary, a crowdsourced, free resource.

Summary

- Performance measured across a variety of languages from different language families.

- This work has influenced similar research in:

  - semantic analysis

  - named-entity mention detection

Thank you

Google™

dipanjand@google.com