

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

Google™

Natural Language Processing in Practice

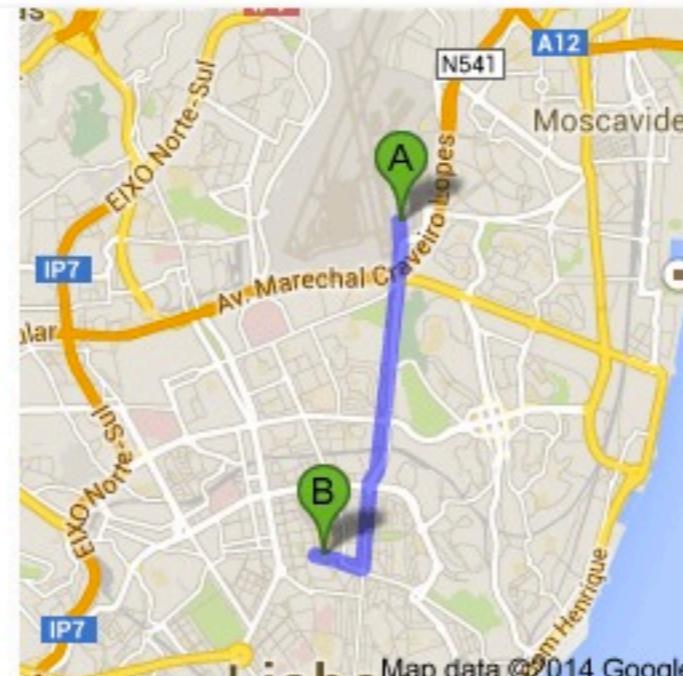
Google directions from lisbon airport to instituto superior tecnico

Web Maps News Images Shopping More Search tools

About 133,000 results (0.57 seconds)

A Lisbon Portela Airport
9 mins
B Instituto Superior Técnico
Avenida Rovisco Pais 1
1049-001 Lisbon, Portugal

2.9 mi



Map data ©2014 Google

Natural Language Processing in Practice

Google™

Google directions from lisbon airport to instituto superior tecnico

Web Maps News Images Shopping More Search tools

About 133,000 results (0.57 seconds)

A Lisbon Portela Airport
9 mins
B Instituto Superior Técnico
Avenida Rovisco Pais 1
1049-001 Lisbon, Portugal

2.9 mi

entity detection

Natural Language Processing in Practice

Google™

resolution
to an action

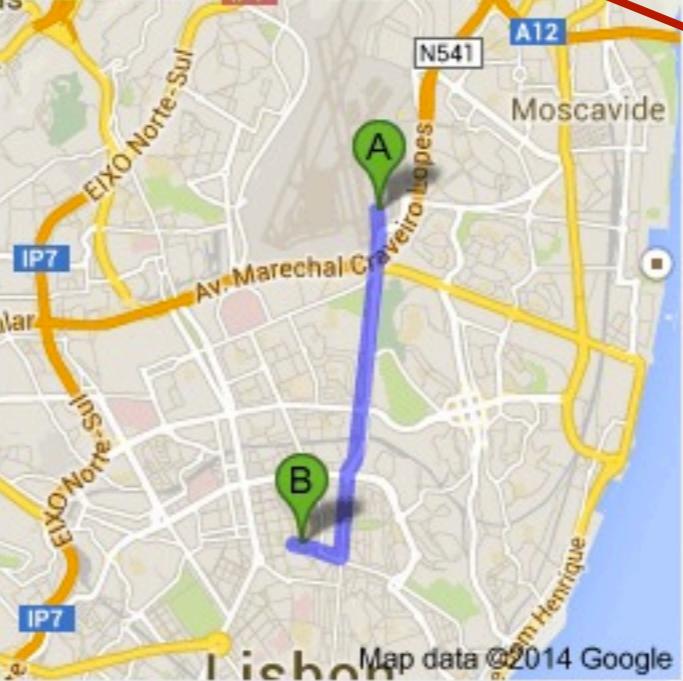
Google 0 Search

Web Maps News Images Shopping More Search tools

About 133,000 results (0.57 seconds)

A Lisbon Portela Airport
9 mins
B Instituto Superior Técnico
Avenida Rovisco Pais 1
1049-001 Lisbon, Portugal

2.9 mi



Map data ©2014 Google

entity detection

Natural Language Processing in Practice Google™

entity linking

Google directions from lisbon airport to instituto superior tecnico

Web Maps News Images Shopping More Search tools

About 133,000 results (0.57 seconds)

A Lisbon Portela Airport
9 mins
B Instituto Superior Técnico
Avenida Rovisco Pais 1
1049-001 Lisbon, Portugal

2.9 mi

Map data ©2014 Google

A B

Natural Language Processing in Practice

 who is the president of the US

[Web](#) [News](#) [Shopping](#) [Images](#) [Maps](#) [More ▾](#) [Search tools](#)

About 3,570,000,000 results (0.54 seconds)

Barack Obama

United States of America, President
President of the United States of America



Natural Language Processing in Practice



Google who is the president of the US ← entity detection

Web News Shopping Images Maps More ▾ Search tools

About 3,570,000,000 results (0.54 seconds)

Barack Obama
United States of America, President
President of the United States of America



Natural Language Processing in Practice



Google who is the president of the US ← entity detection

Web News Shopping Images Maps More ▾ Search tools

About 3,570,000,000 results (0.54 seconds)

Barack Obama
United States of America, President



Natural Language Processing in Practice



Google search interface illustrating NLP components:

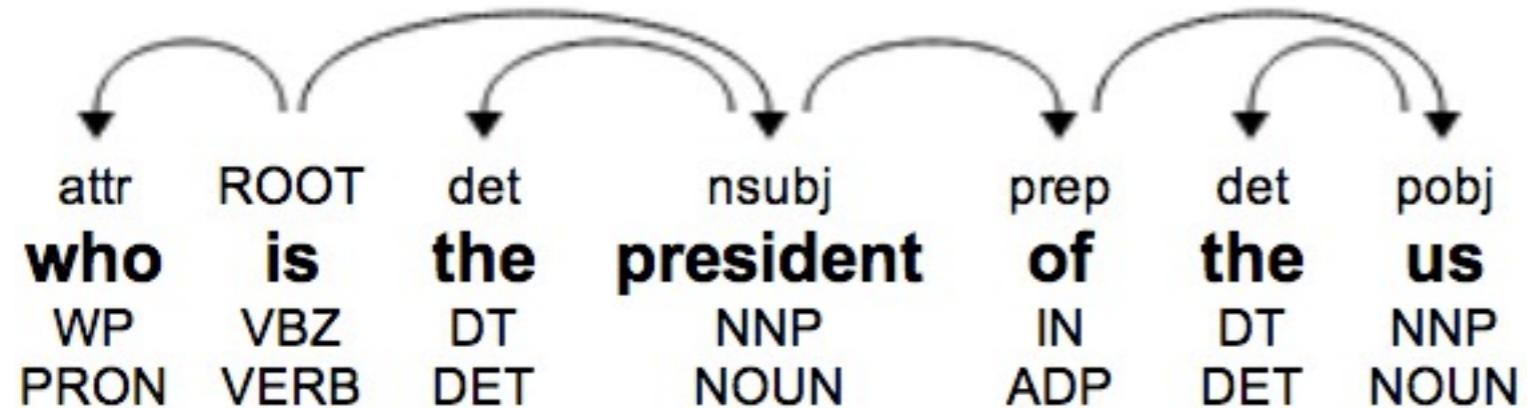
- wh-question**: Points to the search bar containing "who is the president of the US".
- entity detection**: Points to the phrase "the president of the US" which is highlighted with a red oval.
- attribute**: Points to the word "president" which is highlighted with a blue oval.

Search results for "Barack Obama":

Barack Obama
United States of America, President



Natural Language Processing in Practice



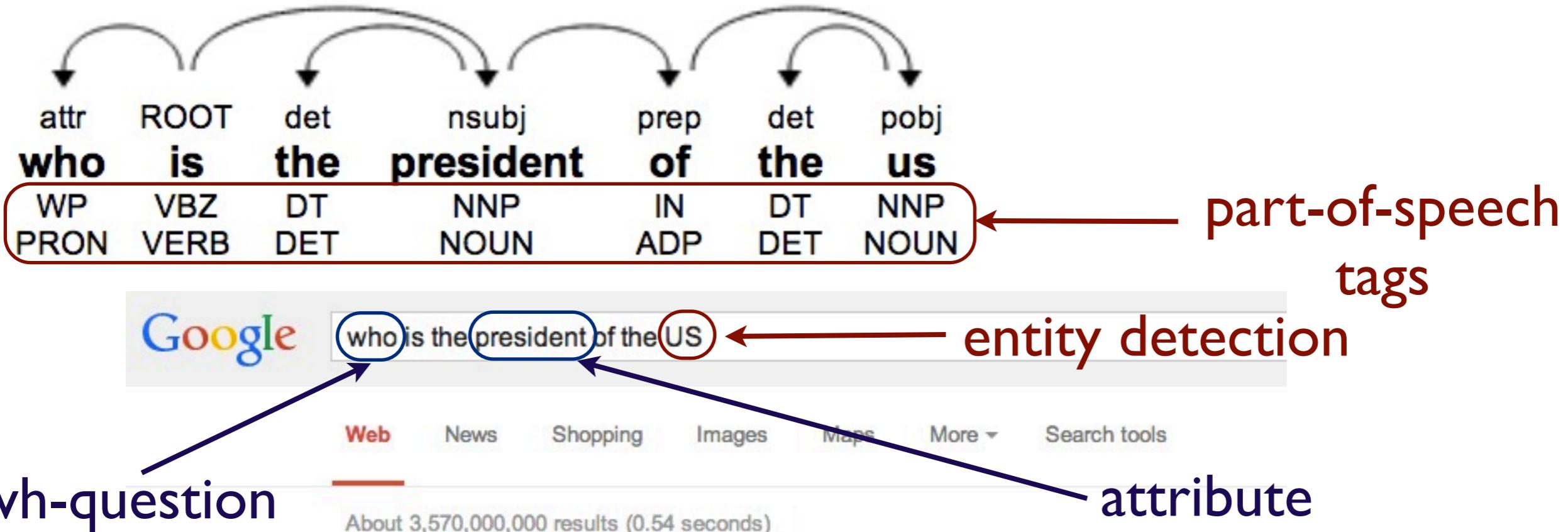
Google search interface showing the query "who is the president of the US".

Annotations:

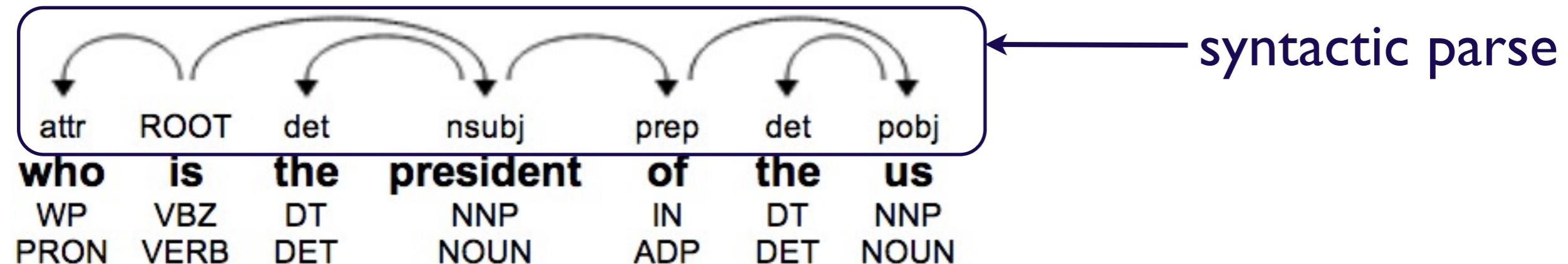
- A blue arrow points from the word "who" in the search bar to the text "wh-question".
- A blue oval highlights the phrase "the president of the US", which is circled in red. A red arrow points from this area to the text "entity detection".
- A dark blue arrow points from the search bar to the text "attribute".



Natural Language Processing in Practice



Natural Language Processing in Practice



entity detection

wh-question

attribute

The screenshot shows a Google search bar with the query "who is the president of the US". The word "who" is highlighted with a blue oval, and the phrase "president of the US" is highlighted with a red oval. Below the search bar, the Google logo is visible. The search results page has a navigation bar with "Web" selected, followed by News, Shopping, Images, Maps, More, and Search tools. A status message at the top says "About 3,570,000,000 results (0.54 seconds)".



Natural Language Processing in Practice

 who is his wife

[Web](#) [News](#) [Images](#) [Videos](#) [Shopping](#) [More ▾](#) [Search tools](#)

About 174,000,000 results (0.93 seconds)

Using previous search [Undo](#)

Michelle Obama (m. 1992)

Barack Obama, Spouse



[Feedback / More info](#)

Natural Language Processing in Practice



who is the president of the US

coreference resolution



who is his wife

Web

News

Images

Videos

Shopping

More ▾

Search tools

About 174,000,000 results (0.93 seconds)

Using previous search [Undo](#)

Michelle Obama (m. 1992)

Barack Obama, Spouse



[Feedback / More info](#)

Natural Language Processing in Practice

Google search results for "who is the president of the US":

who is the **president** of the US

coreference resolution

	NOUN			
who	WP	nssubj	ROOT	
is	VBZ		poss	PRP\$
his	PRON		attr	NN
wife	NOUN			

Google search results for "who is his wife":

who is **his** wife

PRON

Web News Images Videos Shopping More ▾ Search tools

About 174,000,000 results (0.93 seconds)

Using previous search [Undo](#)

Michelle Obama (m. 1992)
Barack Obama, Spouse



Feedback / More info

Natural Language Processing in Practice



coreference resolution

attribute

NOUN

who is the president of the US

PRON

who is his wife

nssubj ROOT poss attr

who is his wife

WP VBZ PRP\$ NN

PRON VERB PRON NOUN

Google

Web News Images Videos Shopping More ▾ Search tools

About 174,000,000 results (0.93 seconds)

Using previous search [Undo](#)

Michelle Obama (m. 1992)

Barack Obama, Spouse



Feedback / More info

Natural Language Processing in Practice

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



who is his wife

coreference
resolution

when was she born

Google

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



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

président des États-Unis

Web Images Vidéos Actualités Maps Plus Outils de recherche

Environ 22 100 000 résultats (0,55 secondes)

Barack Obama

États-Unis, Président



Signaler un problème

Präsident der USA

Web News Bilder Videos Maps Mehr Suchoptionen

Ungefähr 8.960.000 Ergebnisse (0,36 Sekunden)

Barack Obama

Vereinigte Staaten, Präsident



Feedback geben

presidente de los EE.UU

Web Imágenes Noticias Vídeos Maps Más Herramientas de búsqueda

Aproximadamente 70.700.000 resultados (0,56 segundos)

Barack Obama

Estados Unidos, Presidente



Enviar comentarios

президент сша

Поиск Картинки Новости Видео Карты Еще Инструменты поиска

Результатов: примерно 1 580 000 (0,28 сек.)

Обама, Барак

Соединённые Штаты Америки, Президент



Сообщить о проблеме

Goal: 100 languages

Natural Language Processing in Practice

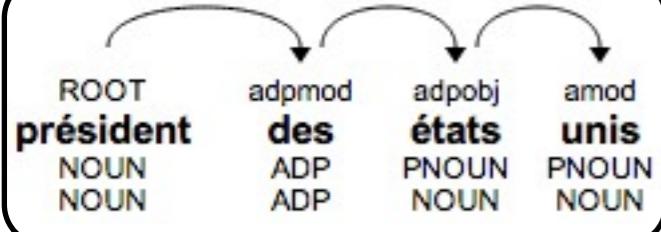
président des États-Unis

Web Images Vidéos Actualités Maps Plus Outils de recherche

Environ 22 100 000 résultats (0,55 secondes)

Barack Obama

États-Unis, Président



Signaler un problème

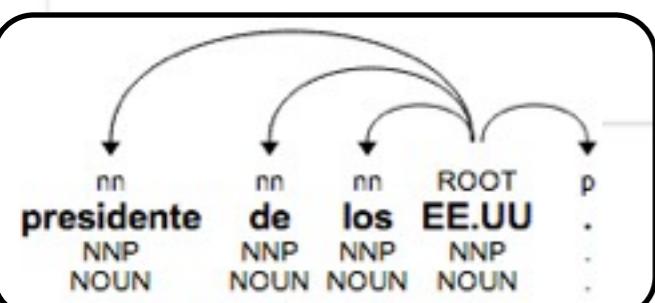
presidente de los EE.UU

Web Imágenes Noticias Vídeos Maps Más Herramientas de búsqueda

Aproximadamente 70.700.000 resultados (0,56 segundos)

Barack Obama

Estados Unidos, Presidente



Enviar comentarios

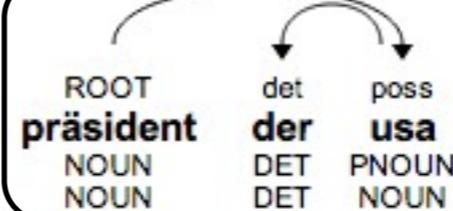
Präsident der USA

Web News Bilder Videos Maps Mehr Suchoptionen

Ungefähr 8.960.000 Ergebnisse (0,36 Sekunden)

Barack Obama

Vereinigte Staaten, Präsident



Feedback geben

президент сша

Поиск Картинки Новости Видео Карты Ещё Инструменты поиска

Результатов: примерно 1 580 000 (0,28 сек.)

Обама, Барак

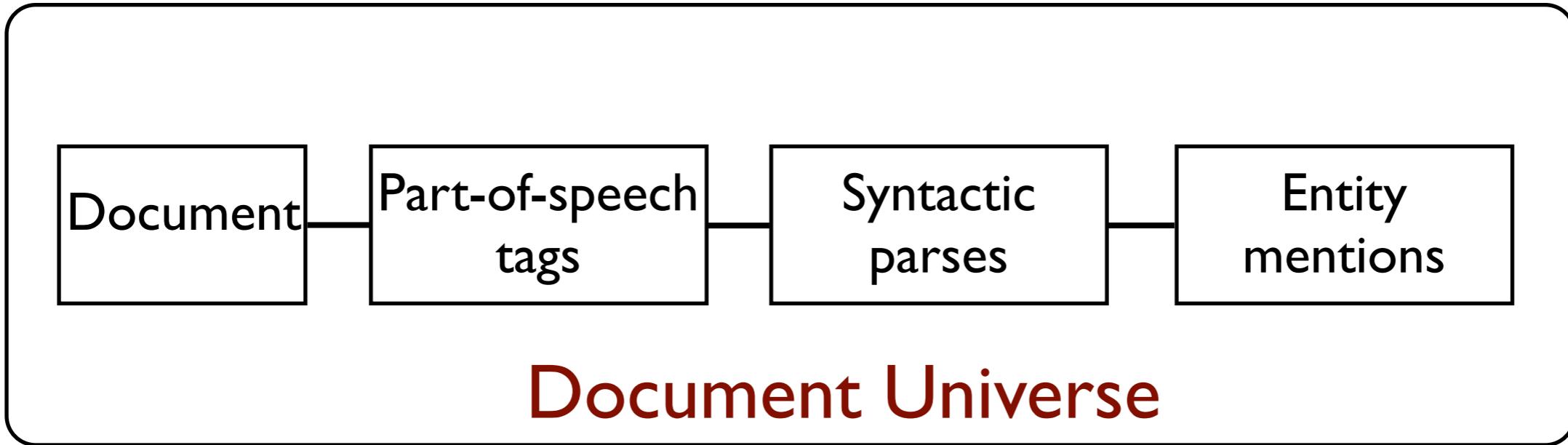
Соединённые Штаты Америки, Президент



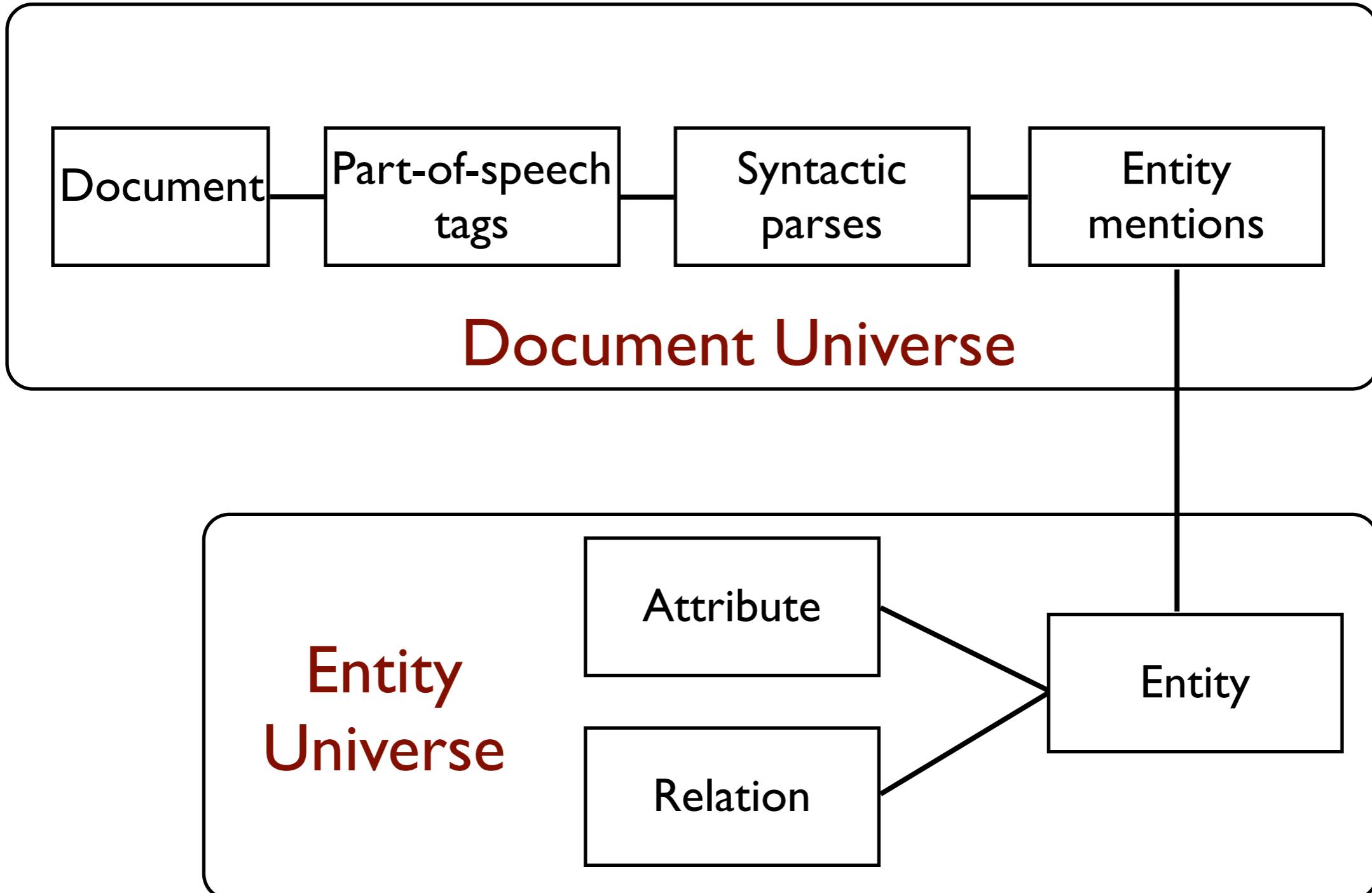
Сообщить о проблеме

Goal: 100 languages

Natural Language Processing in Practice



Natural Language Processing in Practice



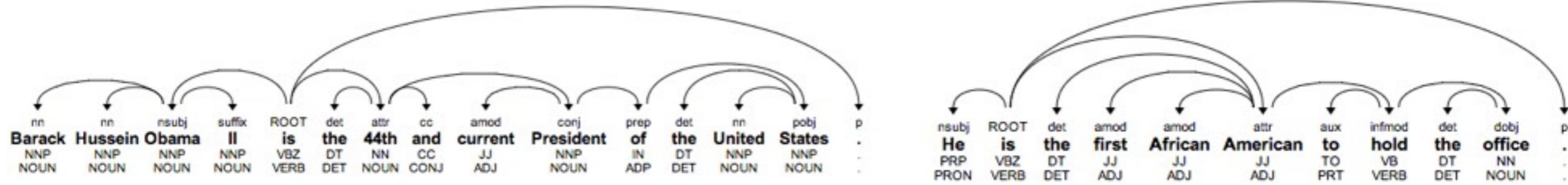
Natural Language Processing in Practice



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

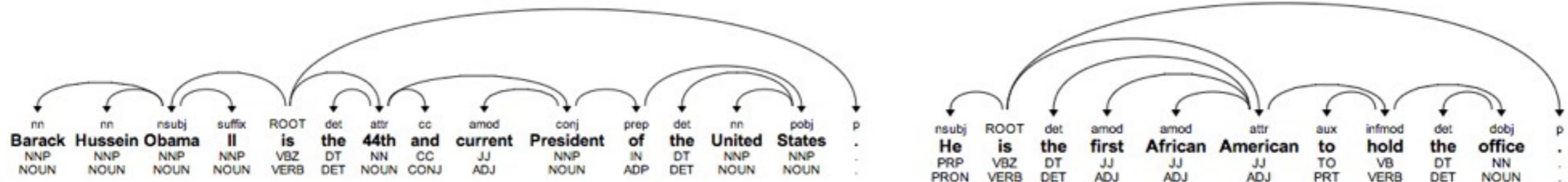
Natural Language Processing in Practice

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

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

E₁ Barack Hussein Obama II

PER

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

E
C
T

E₂ President

PER

Mentions	President
----------	-----------

E₃ United States

LOC

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

E₄ African American

ORG

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

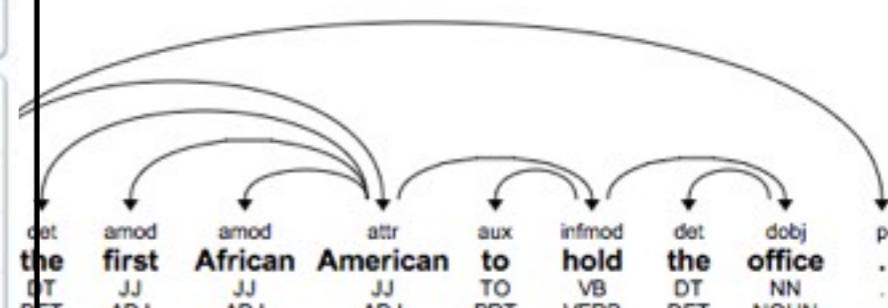
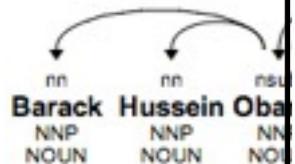
E₅ office

ORG

Mentions	office
----------	--------

Practice Google™

he 44th and States. He is 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.

NLP Pipeline



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

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

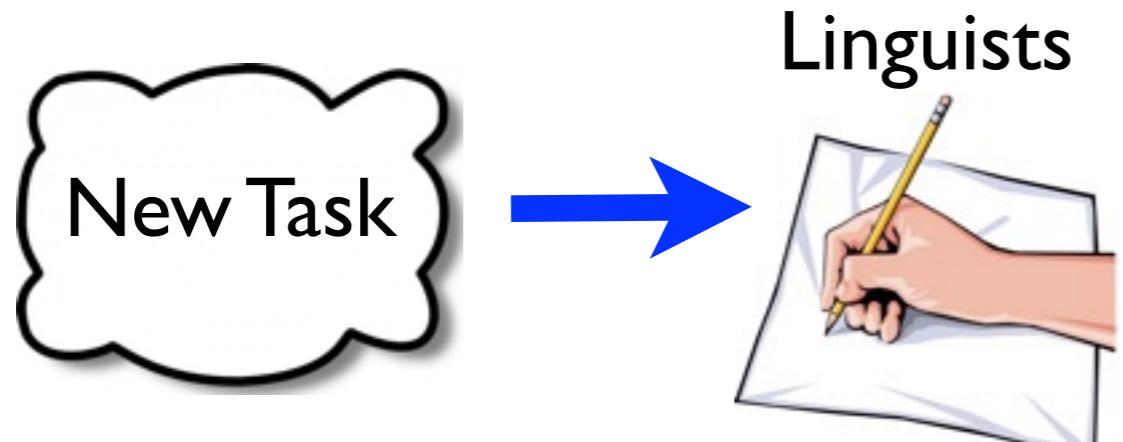
- Build accurate natural language systems
 - Part-of-speech tagging
 - Parsing
 - Entity mention detection
 - Entity Linking
 - Relation Extraction
 - ...
-
- Goal: Build such systems in ~100 languages

Syntax

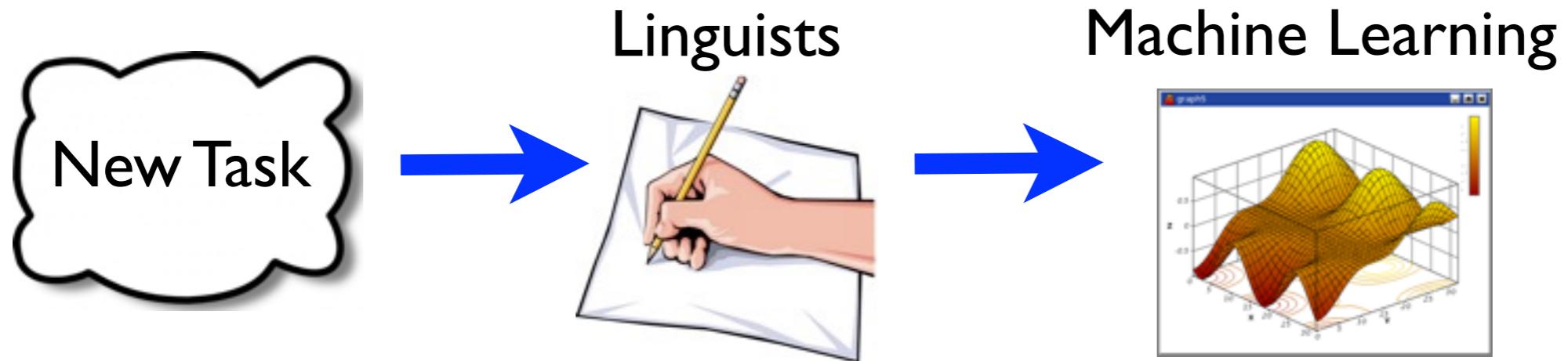
Traditional Supervised NLP



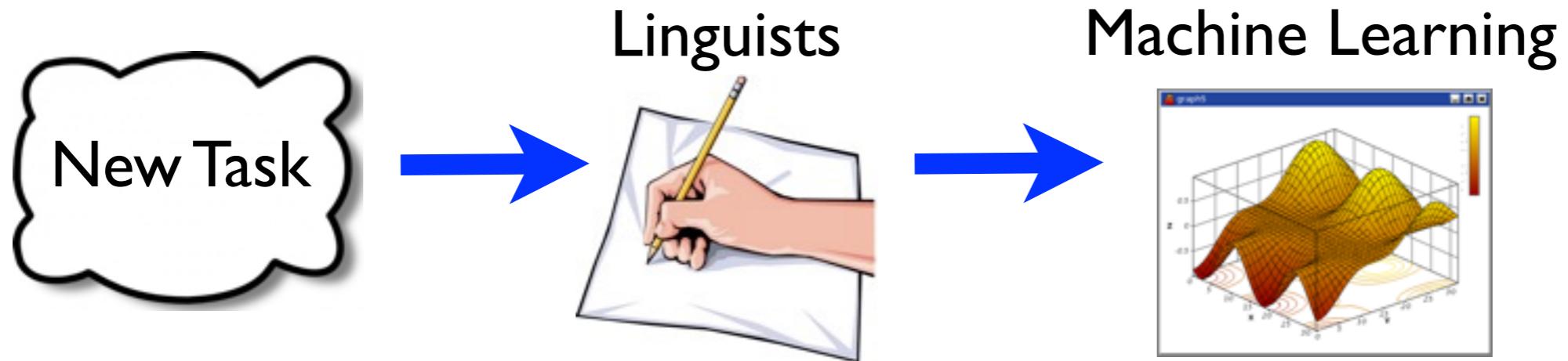
Traditional Supervised NLP



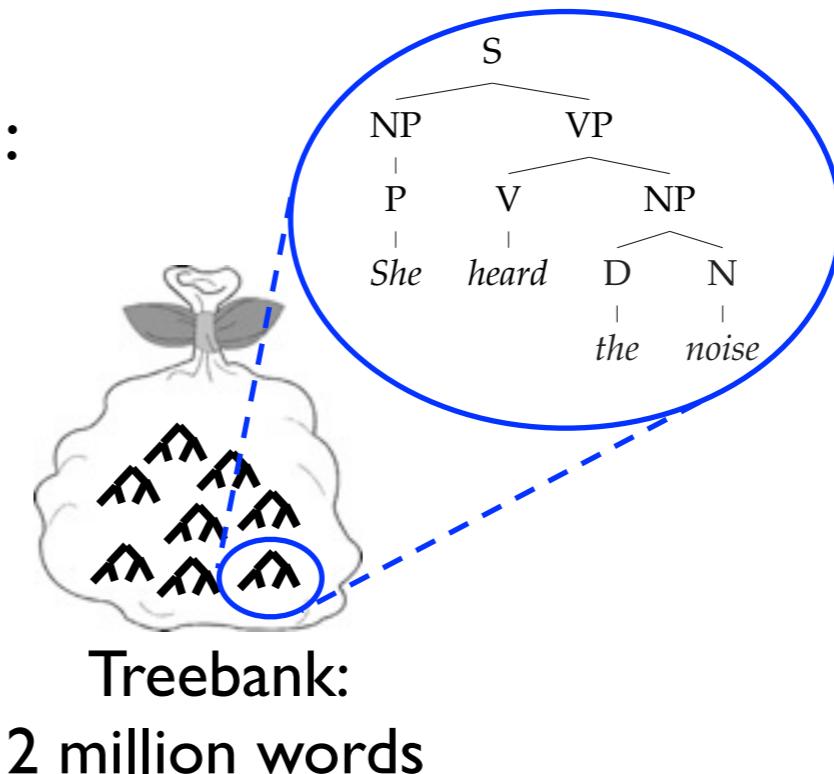
Traditional Supervised NLP



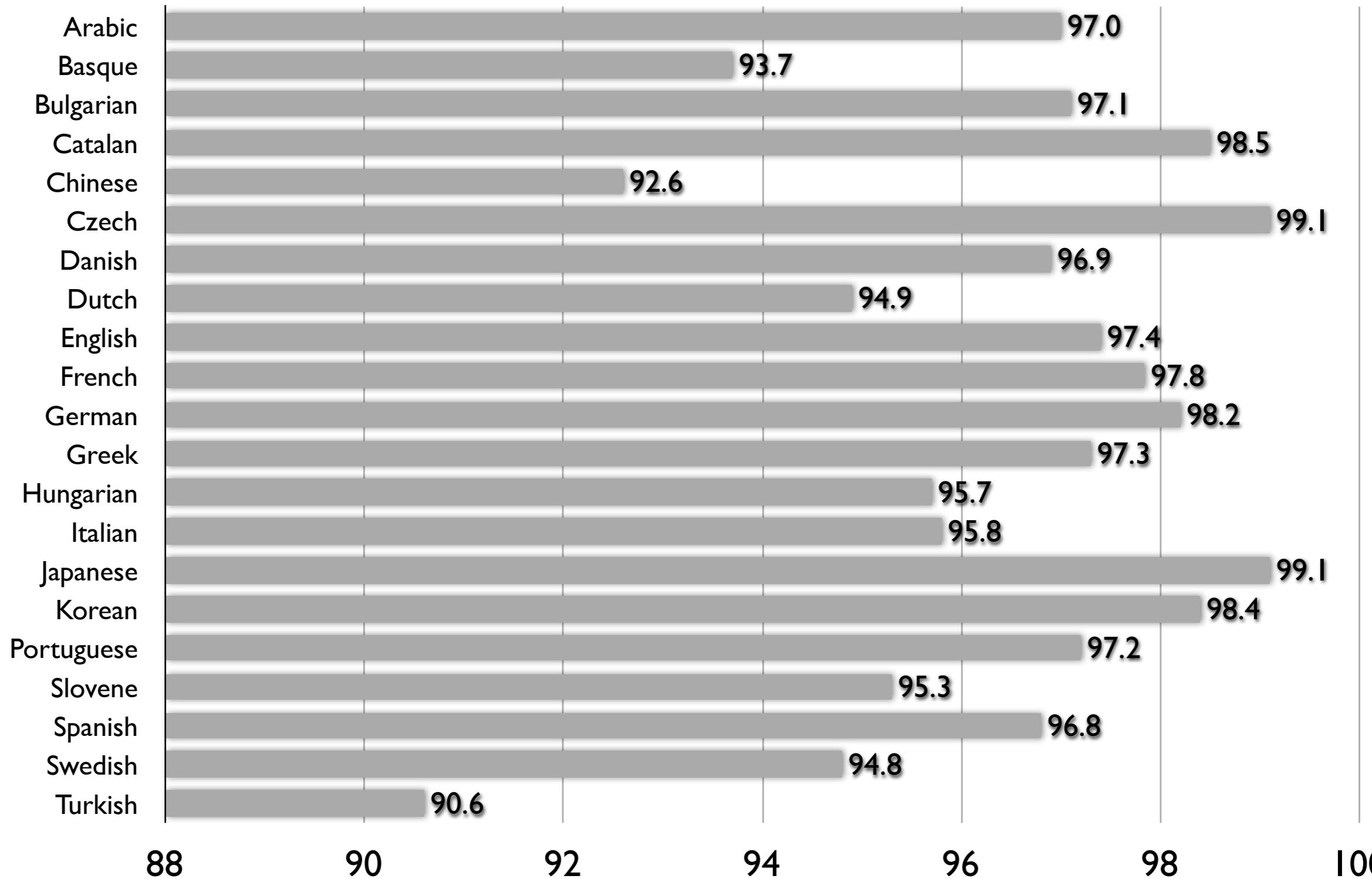
Traditional Supervised NLP



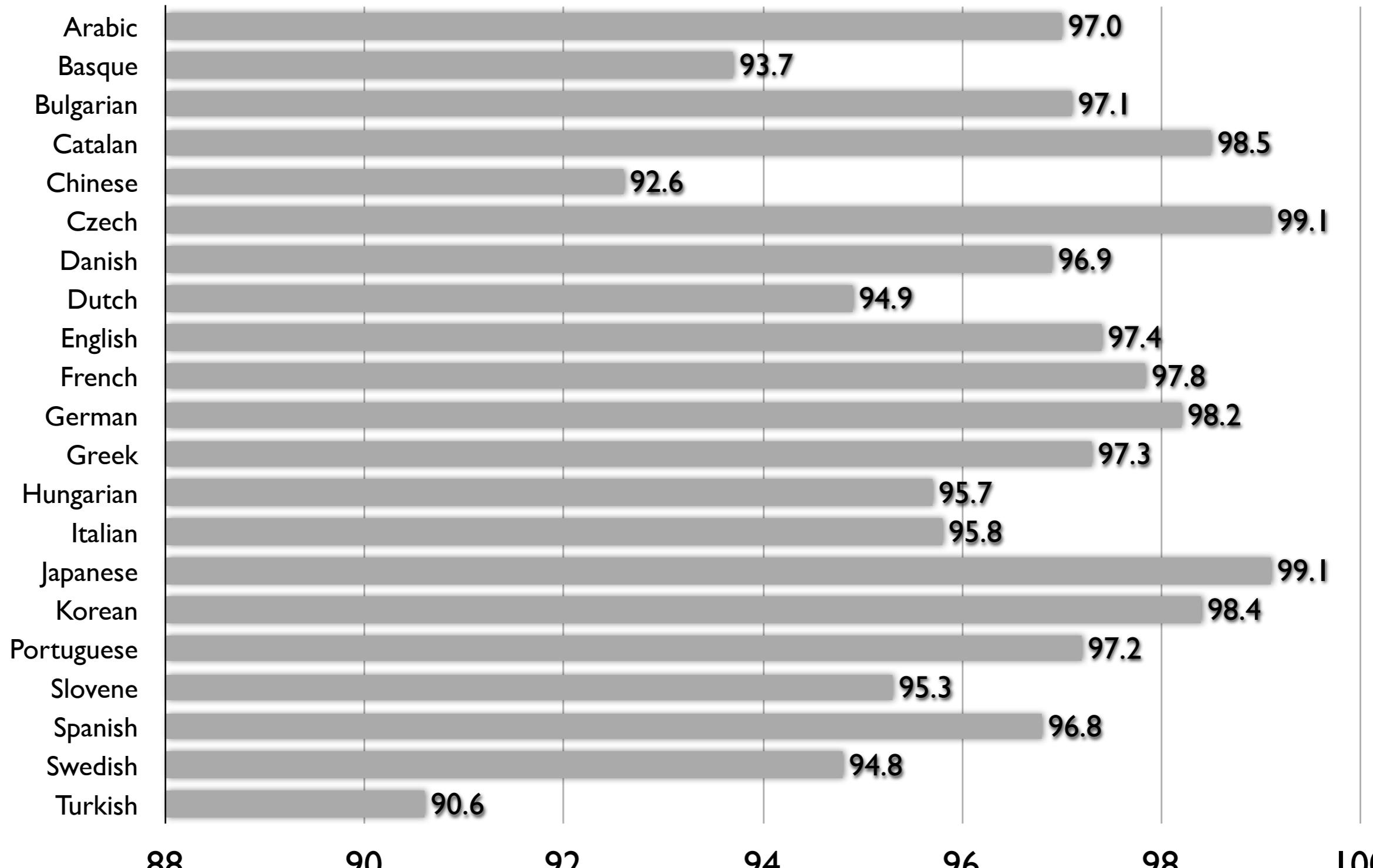
- Example: Syntactic analysis for English:
 - Not solved, but accuracies are high
 - 97% for parts-of-speech
 - 93% for parse trees



Supervised Part-of-Speech Tagging



Supervised Part-of-Speech Tagging



Average accuracy is 96.2% (Brants, 2000)

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

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

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



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

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

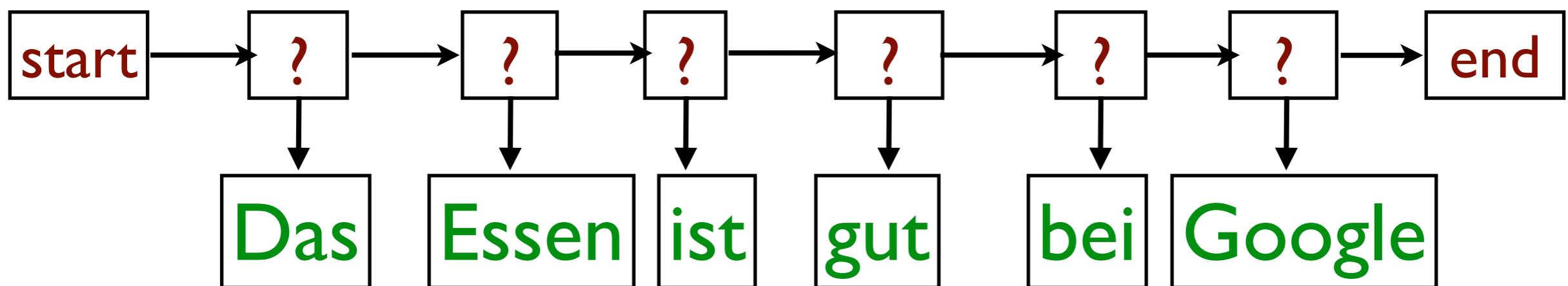
Google

Merialdo (1994)

Unsupervised Part-of-Speech Tagging



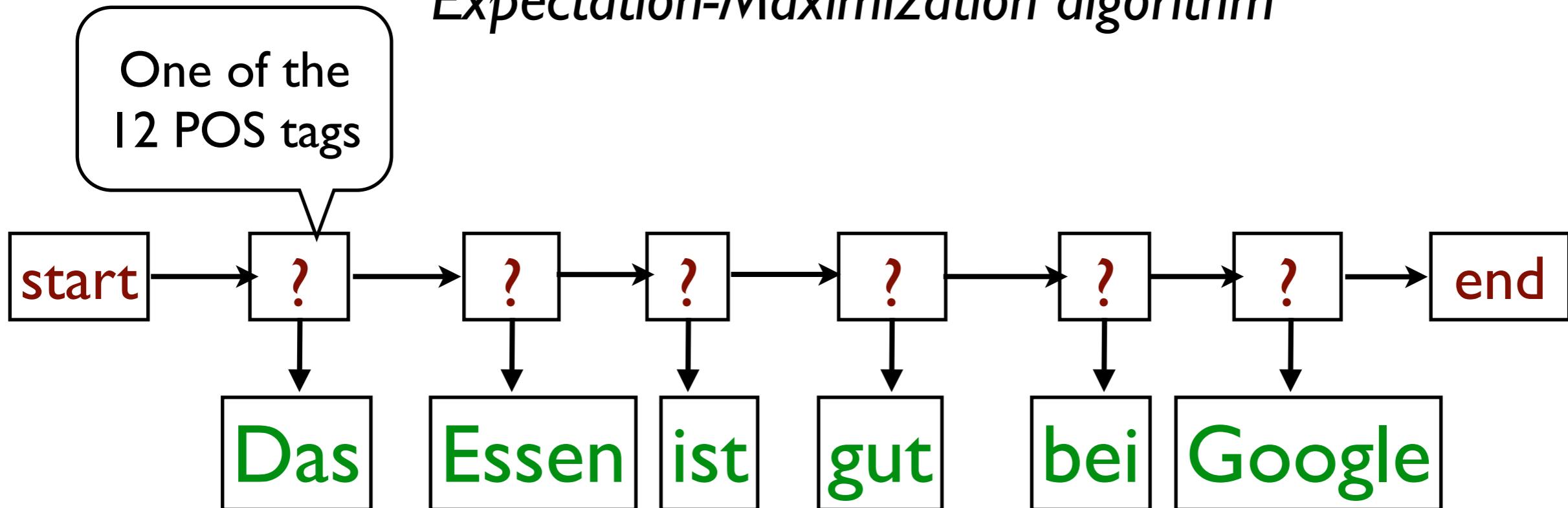
*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



Unsupervised Part-of-Speech Tagging

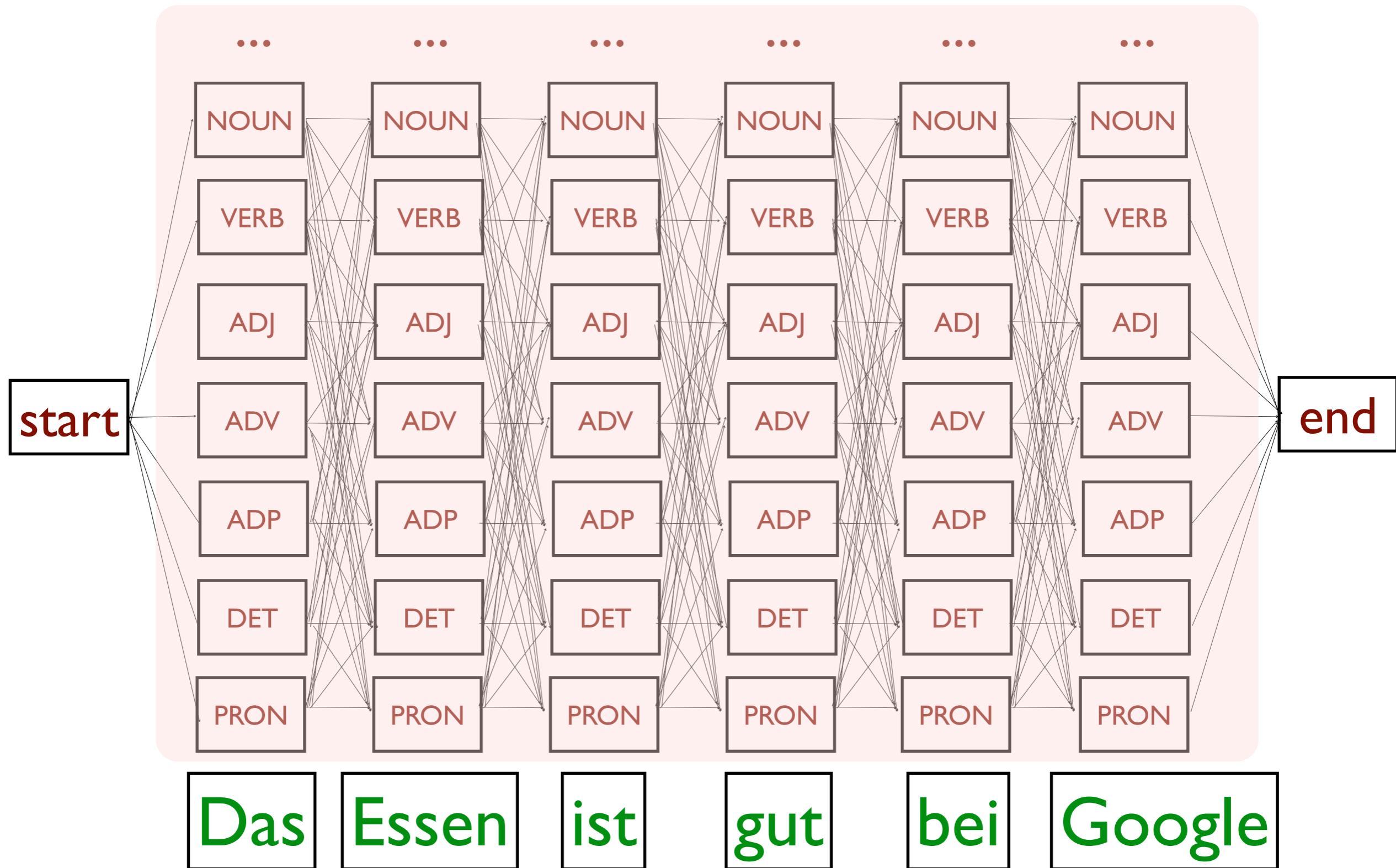


Das Essen ist gut bei Google

Unsupervised Part-of-Speech Tagging



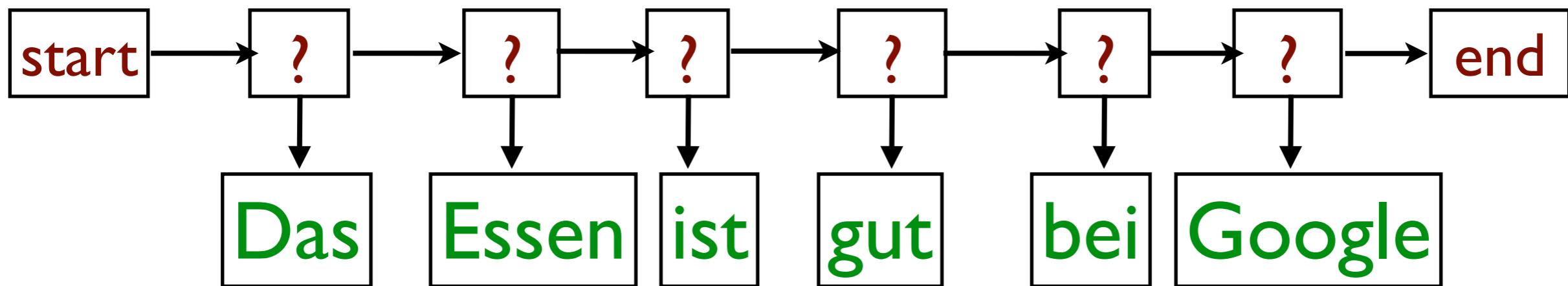
HMM lattice



Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



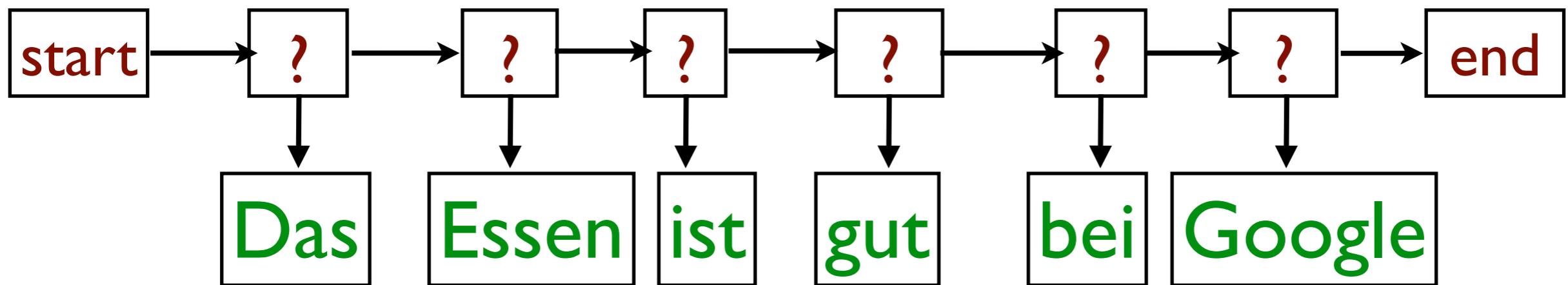
x : word sequence

y : tag sequence

Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



x : word sequence

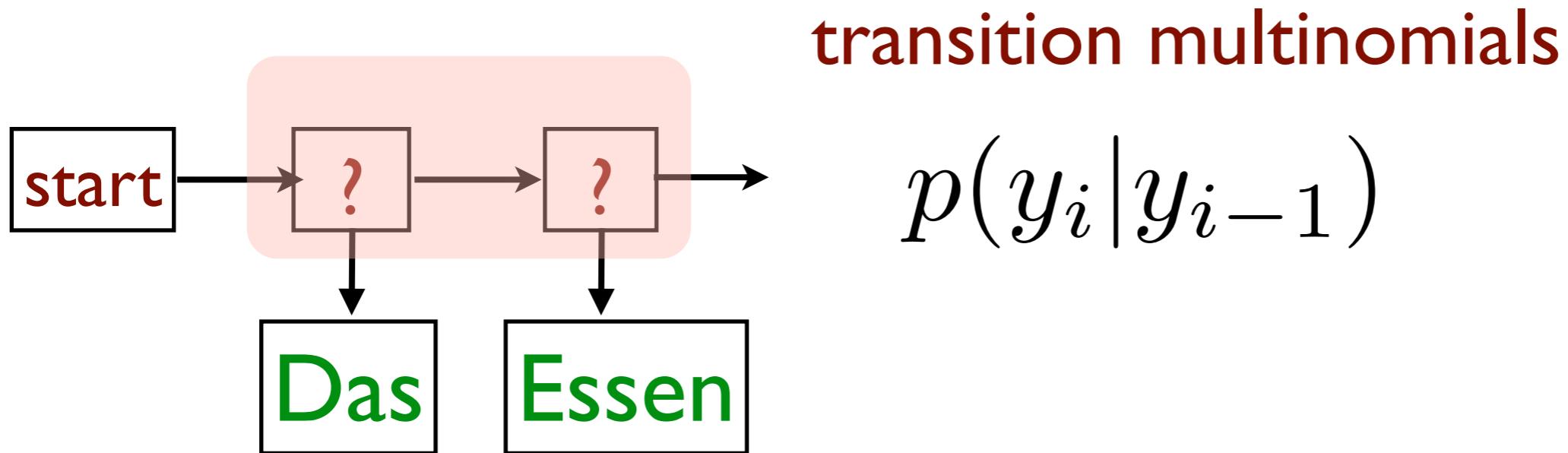
y : tag sequence

Model
 $p(x, y)$

Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



x : word sequence

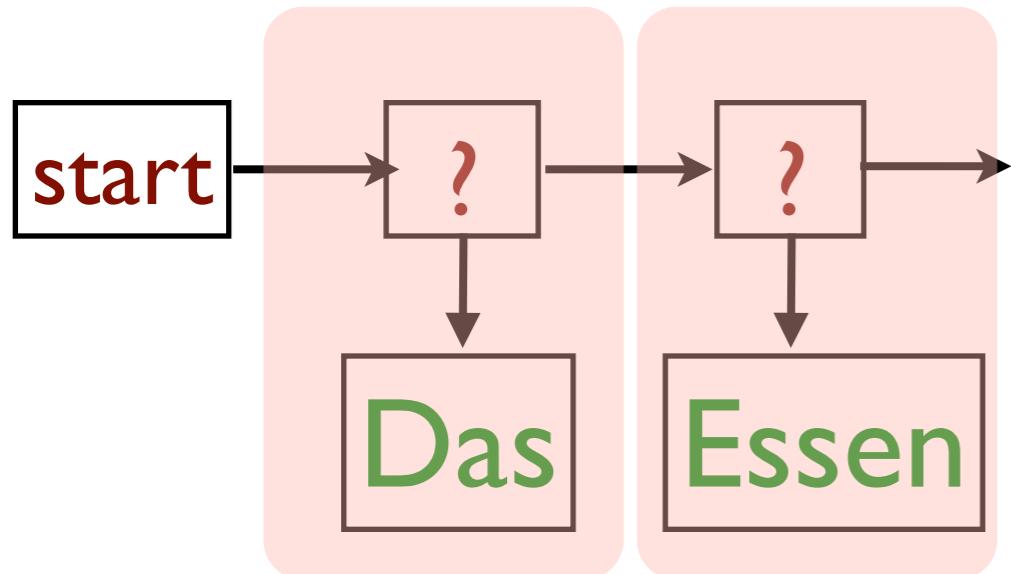
y : tag sequence

Model
 $p(x, y)$

Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



emission multinomials

$$p(x_i | y_i)$$

x : word sequence

y : tag sequence

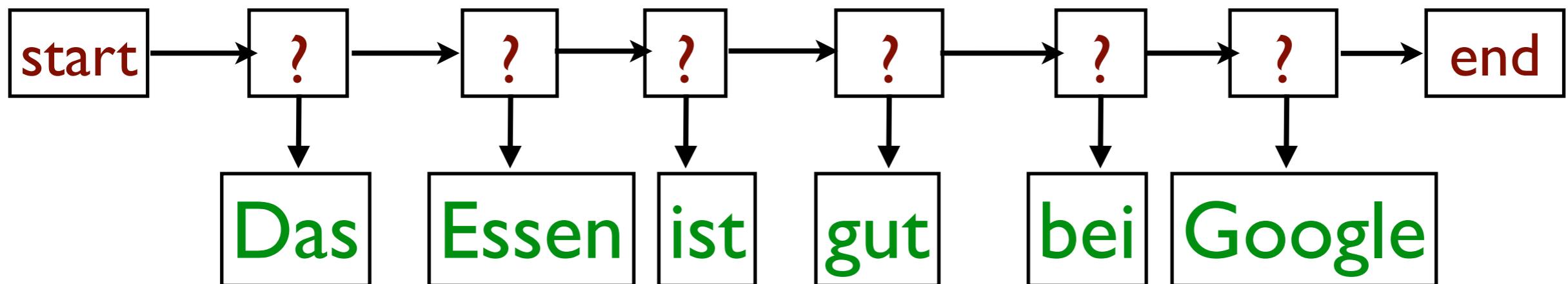
Model

$$p(x, y)$$

Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*

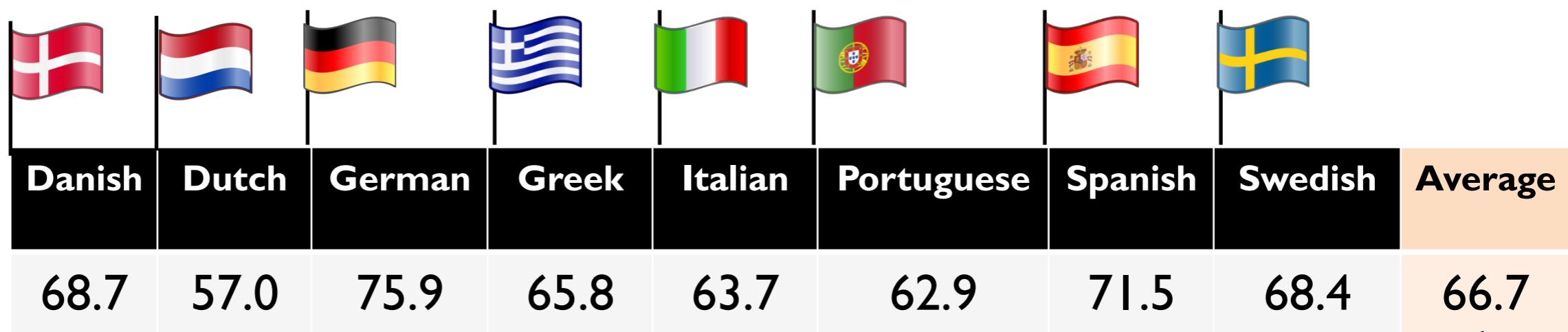
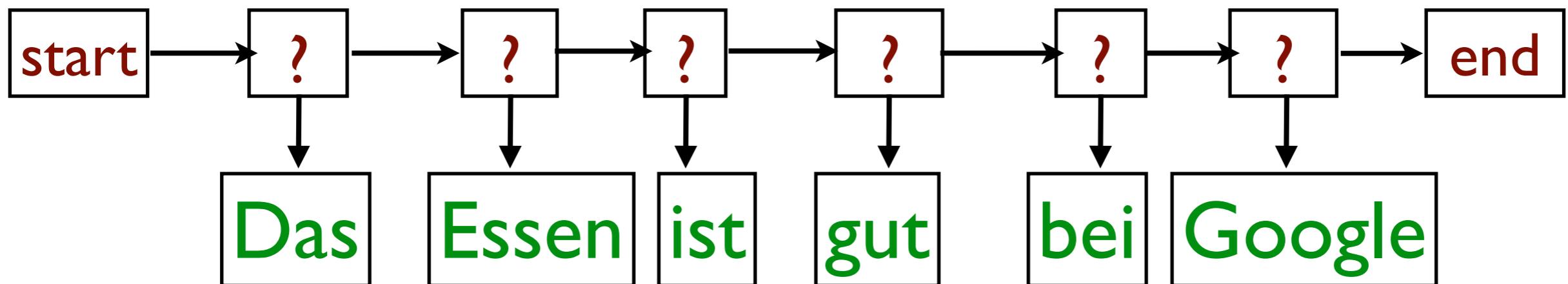


									Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7

Unsupervised Part-of-Speech Tagging



*Hidden Markov Model estimated with the
Expectation-Maximization algorithm*



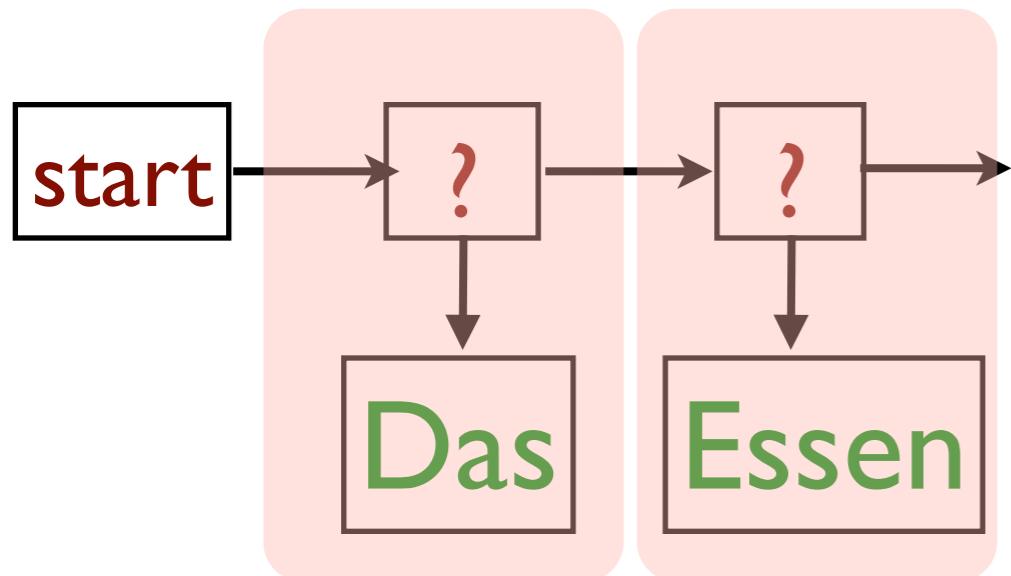
Johnson (2007)

Poor average
result

Unsupervised Part-of-Speech Tagging



HMM with locally-normalized log-linear models



emission multinomials

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

Berg-Kirkpatrick et al. (2010)

x : word sequence

y : tag sequence

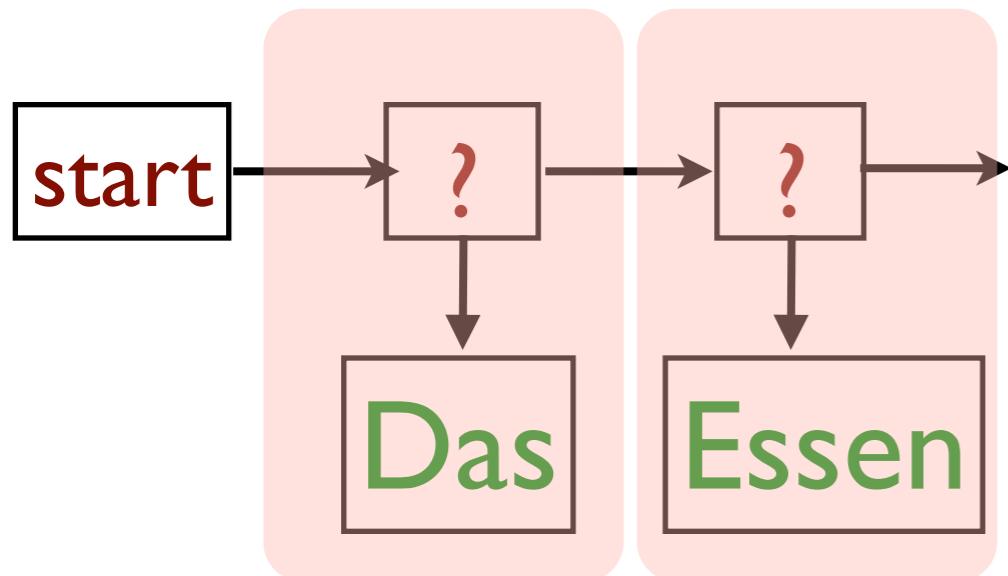
Model

$$p(x, y)$$

Unsupervised Part-of-Speech Tagging



HMM with locally-normalized log-linear models



emission multinomials

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

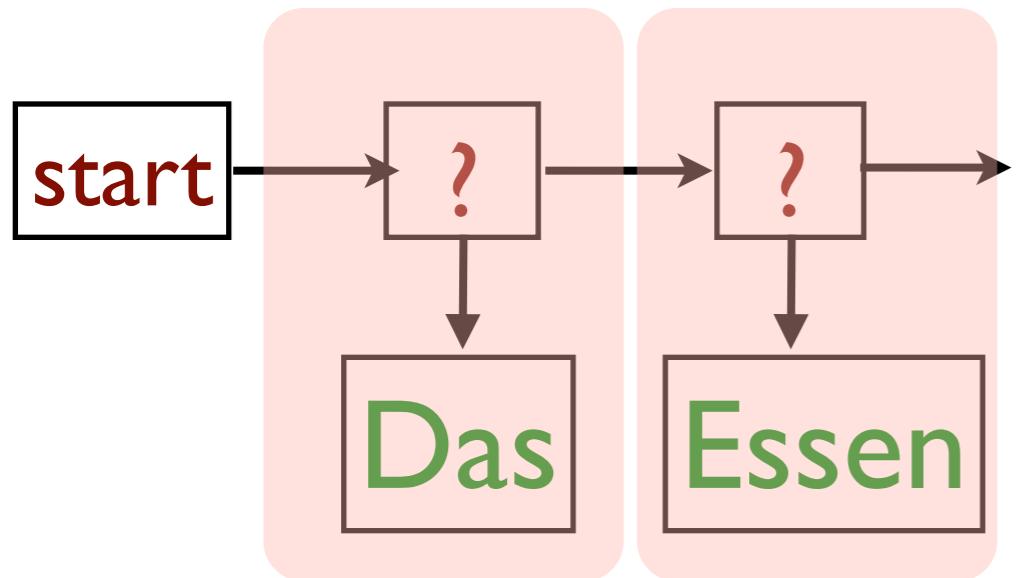
x : word sequence
 y : tag sequence

suffix
hyphen
capitalization
numbers

Unsupervised Part-of-Speech Tagging

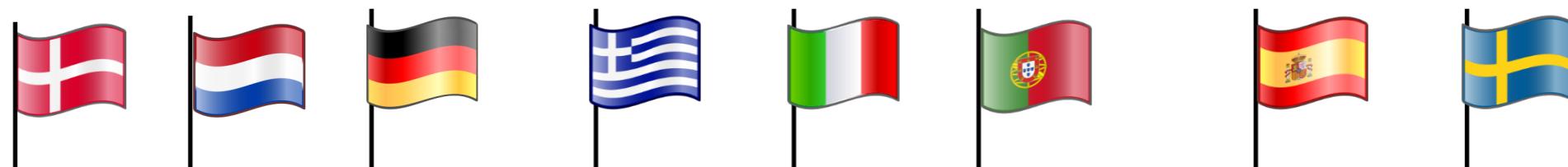


HMM with locally-normalized log-linear models



emission multinomials

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

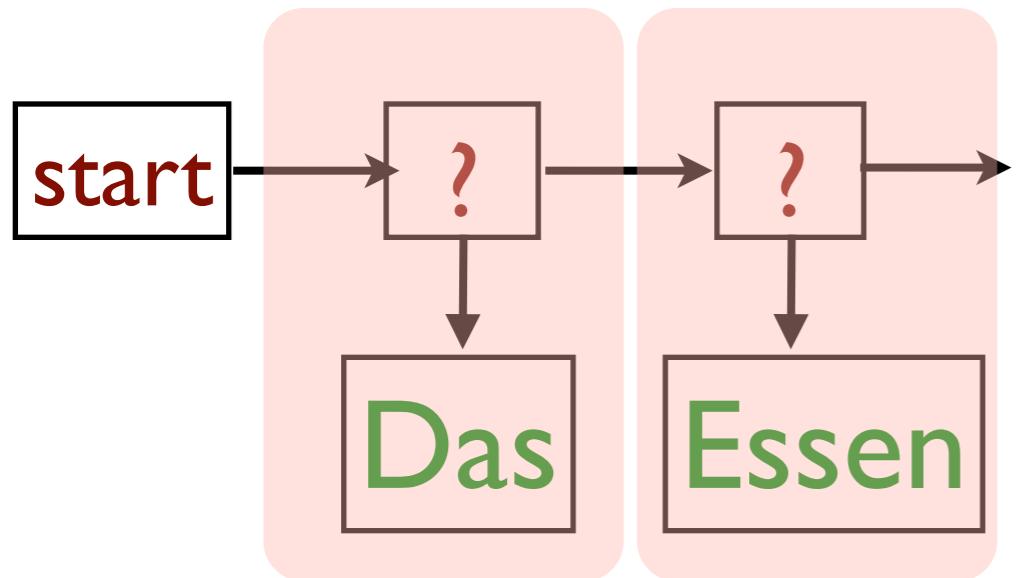


	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0

Unsupervised Part-of-Speech Tagging

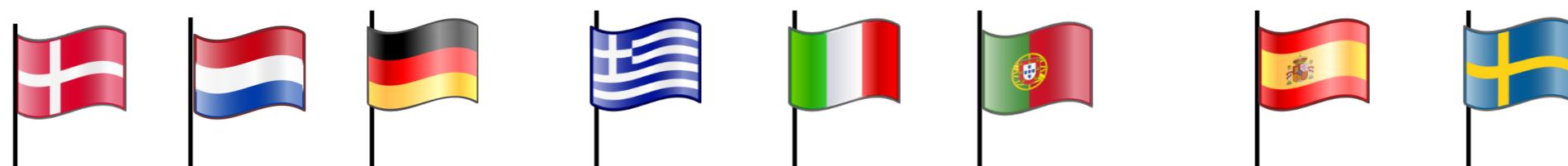


HMM with locally-normalized log-linear models



emission multinomials

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



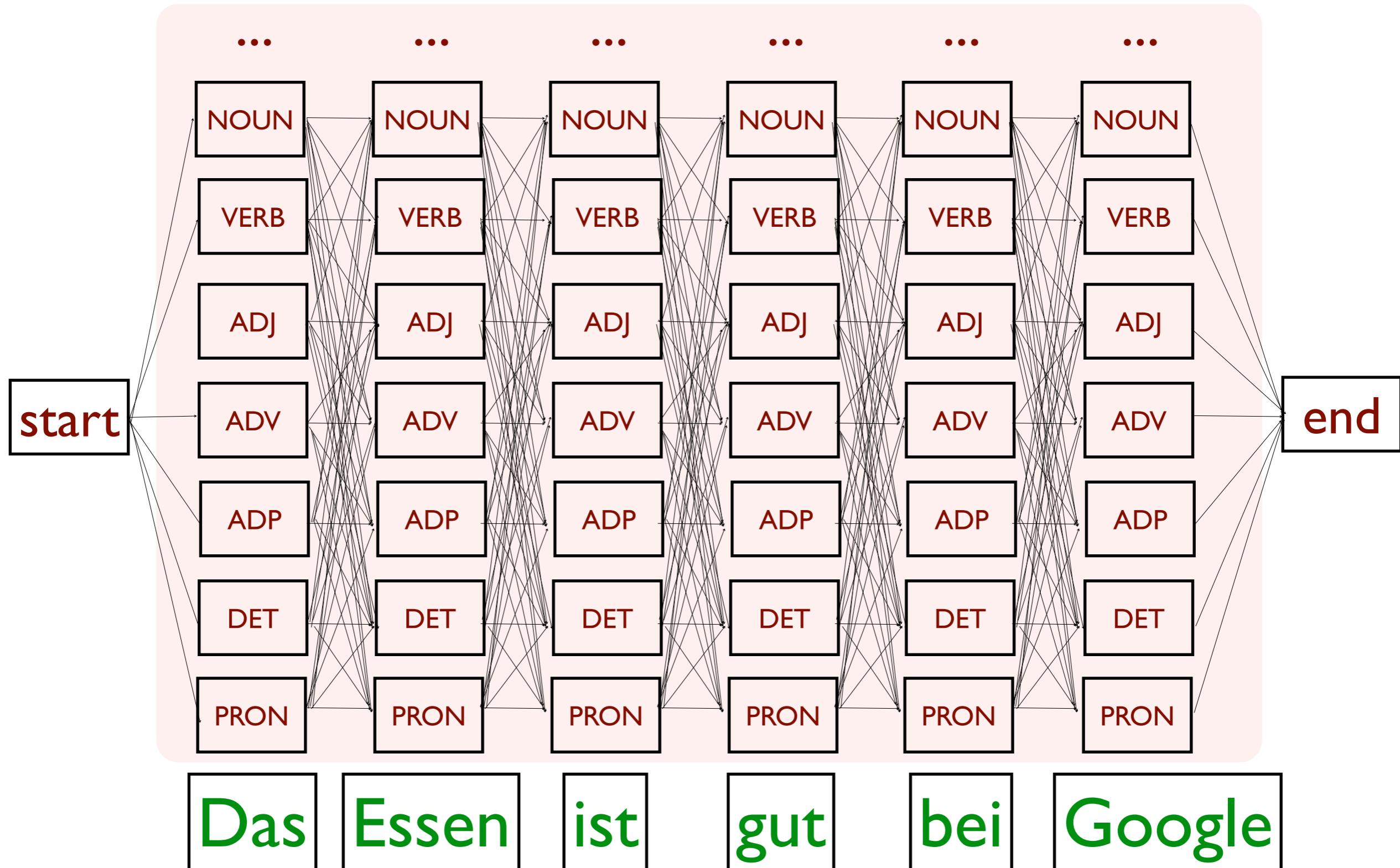
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0

Improvements across all languages

Unsupervised POS Tagging with Dictionaries



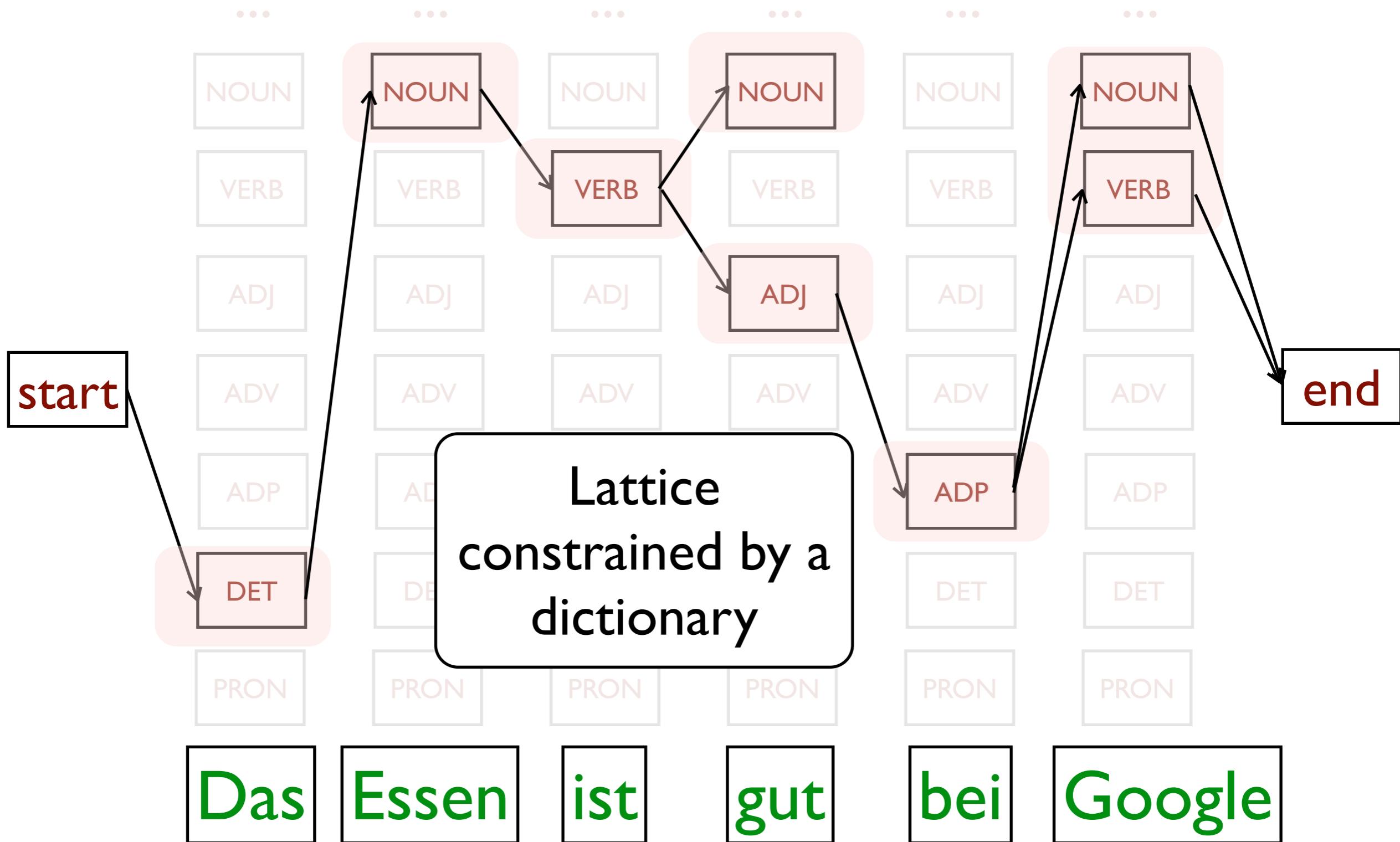
HMM with locally-normalized log-linear models



Unsupervised POS Tagging with Dictionaries



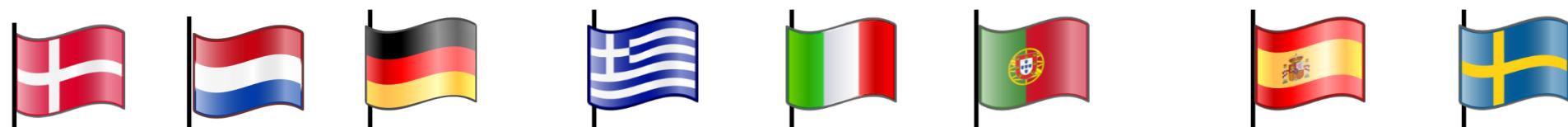
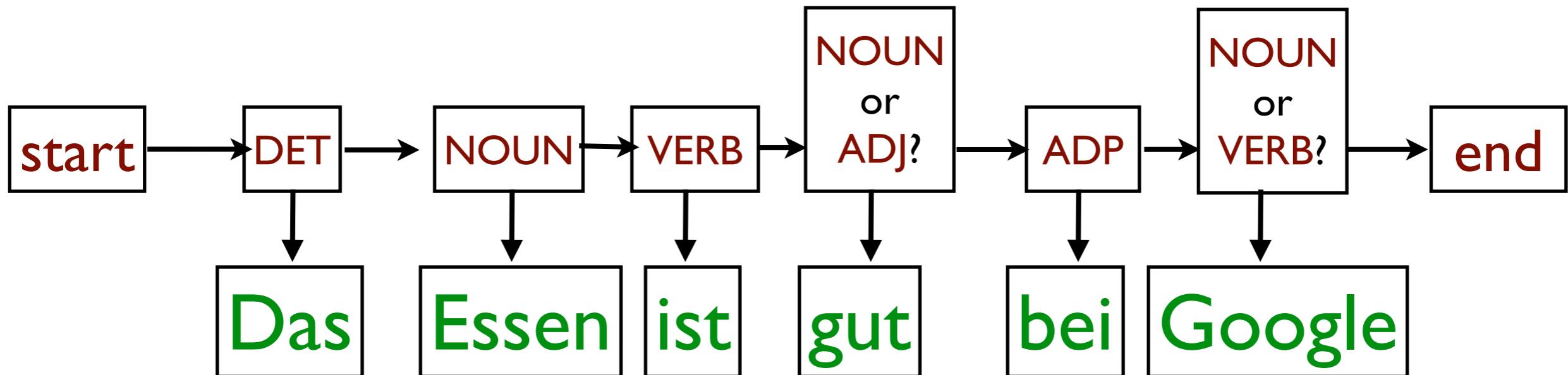
HMM with *locally-normalized log-linear models*



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

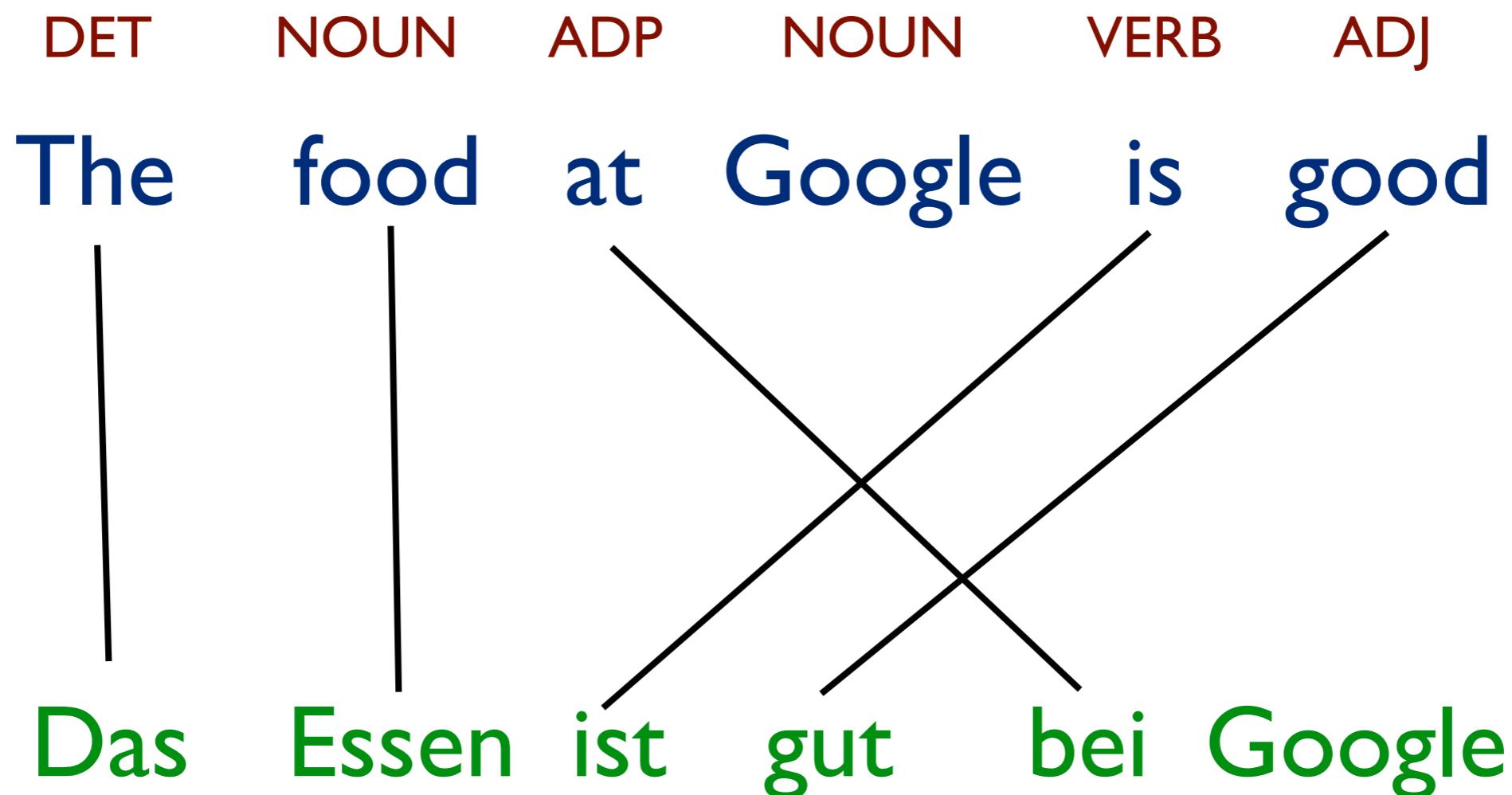


DET NOUN ADP NOUN VERB ADJ

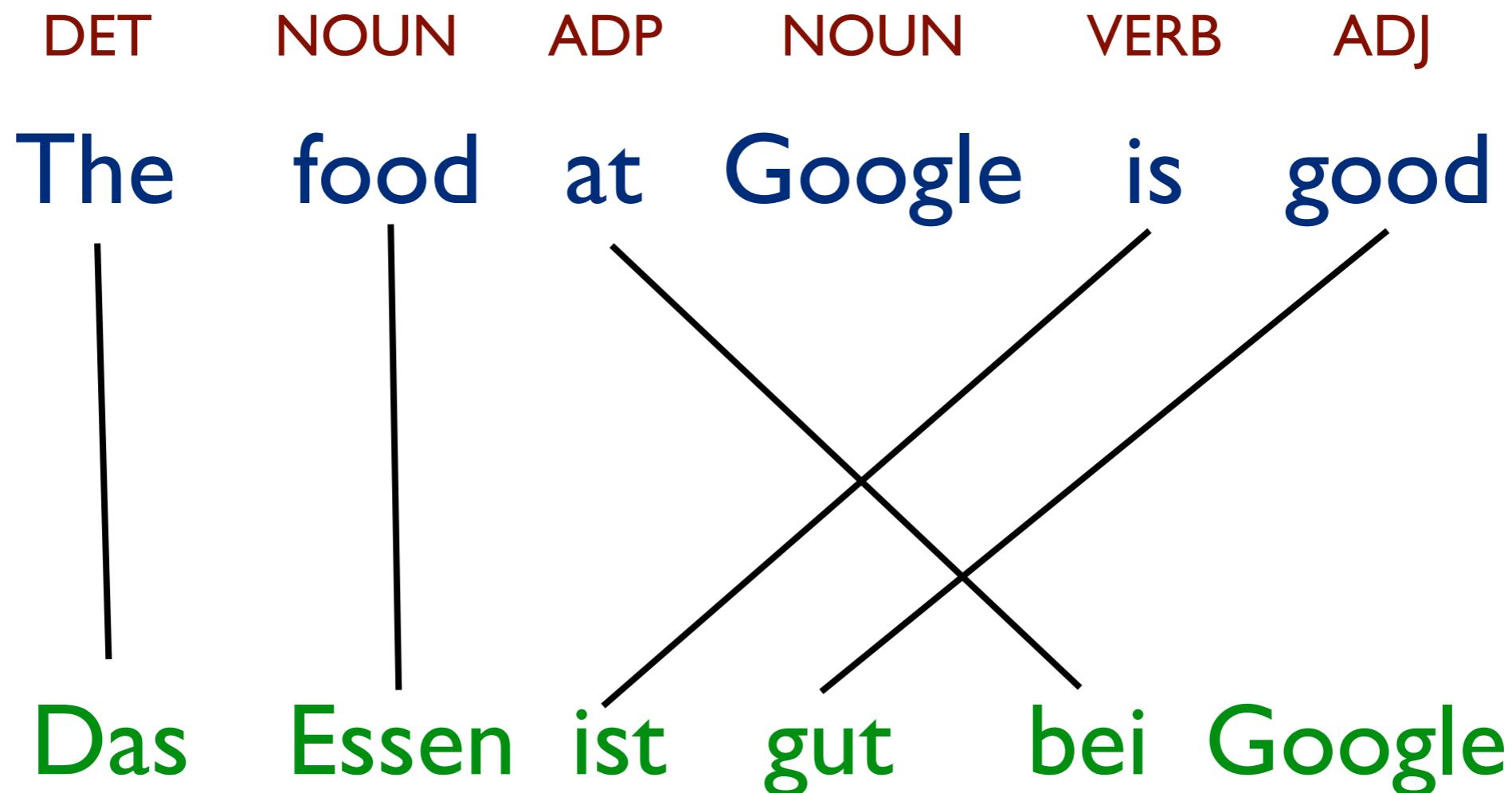
The food at Google is good

Das Essen ist gut bei Google

Cross-Lingual Projection

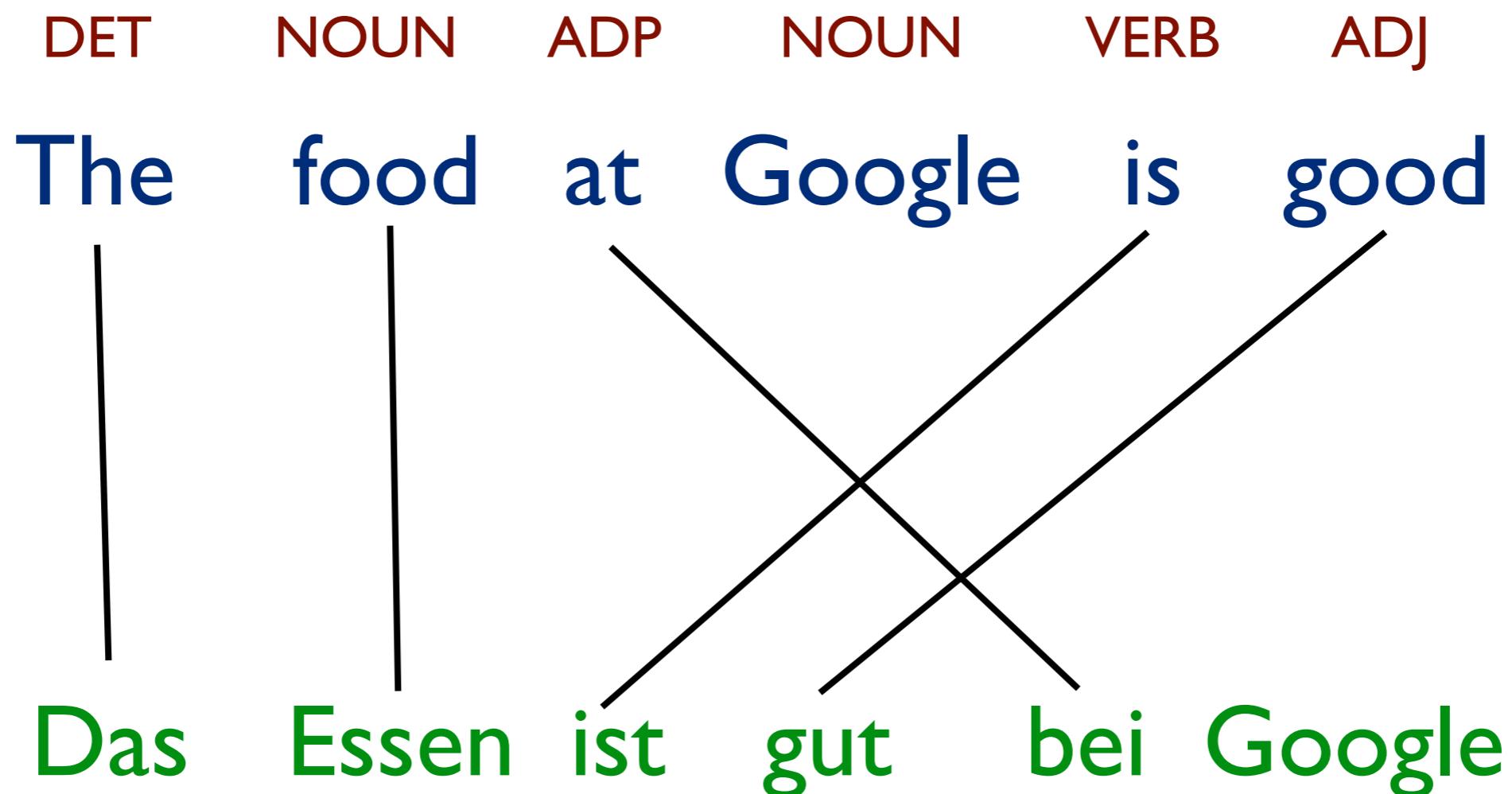


Cross-Lingual Projection

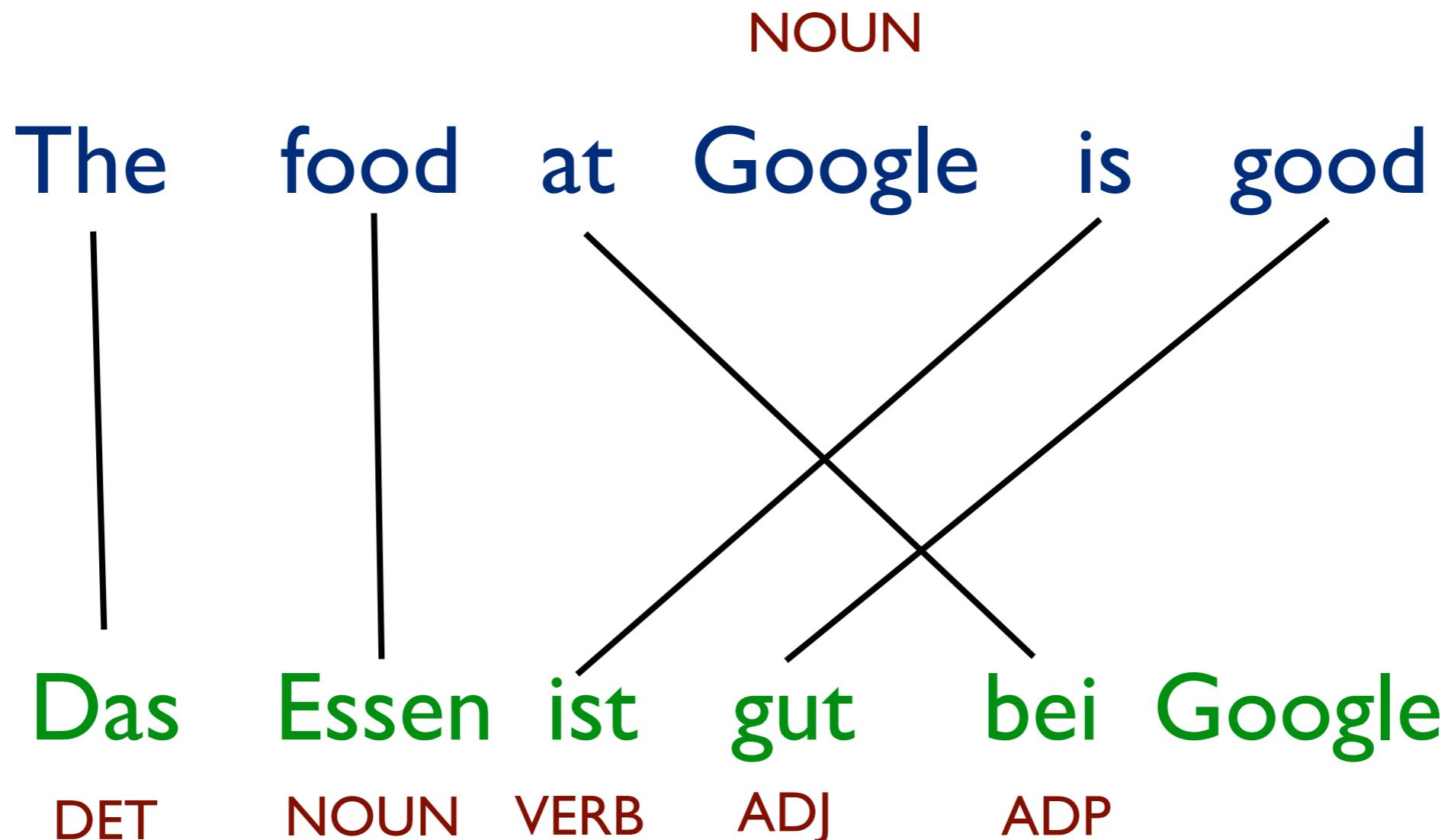


Automatic unsupervised alignments from translation data
(available for ~100 languages)

Cross-Lingual Projection



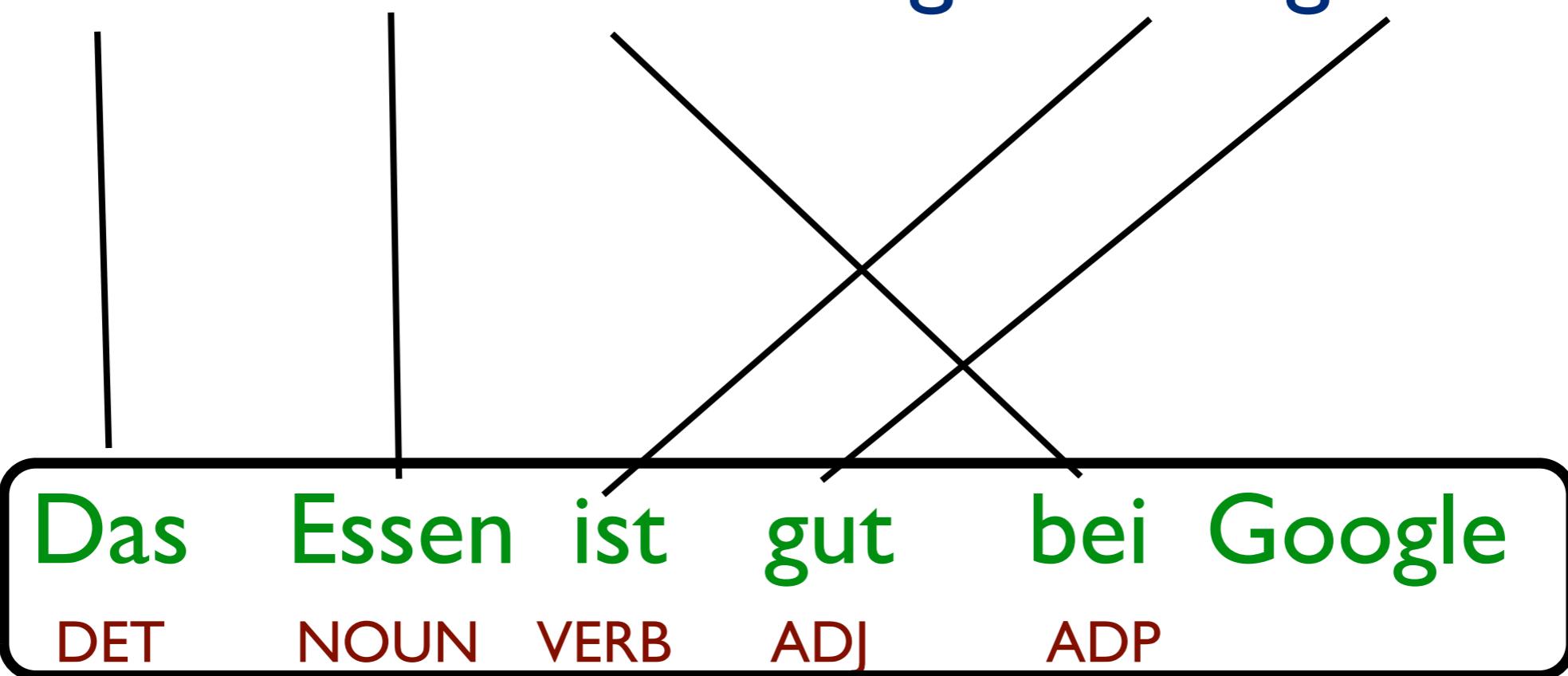
Cross-Lingual Projection



Cross-Lingual Projection

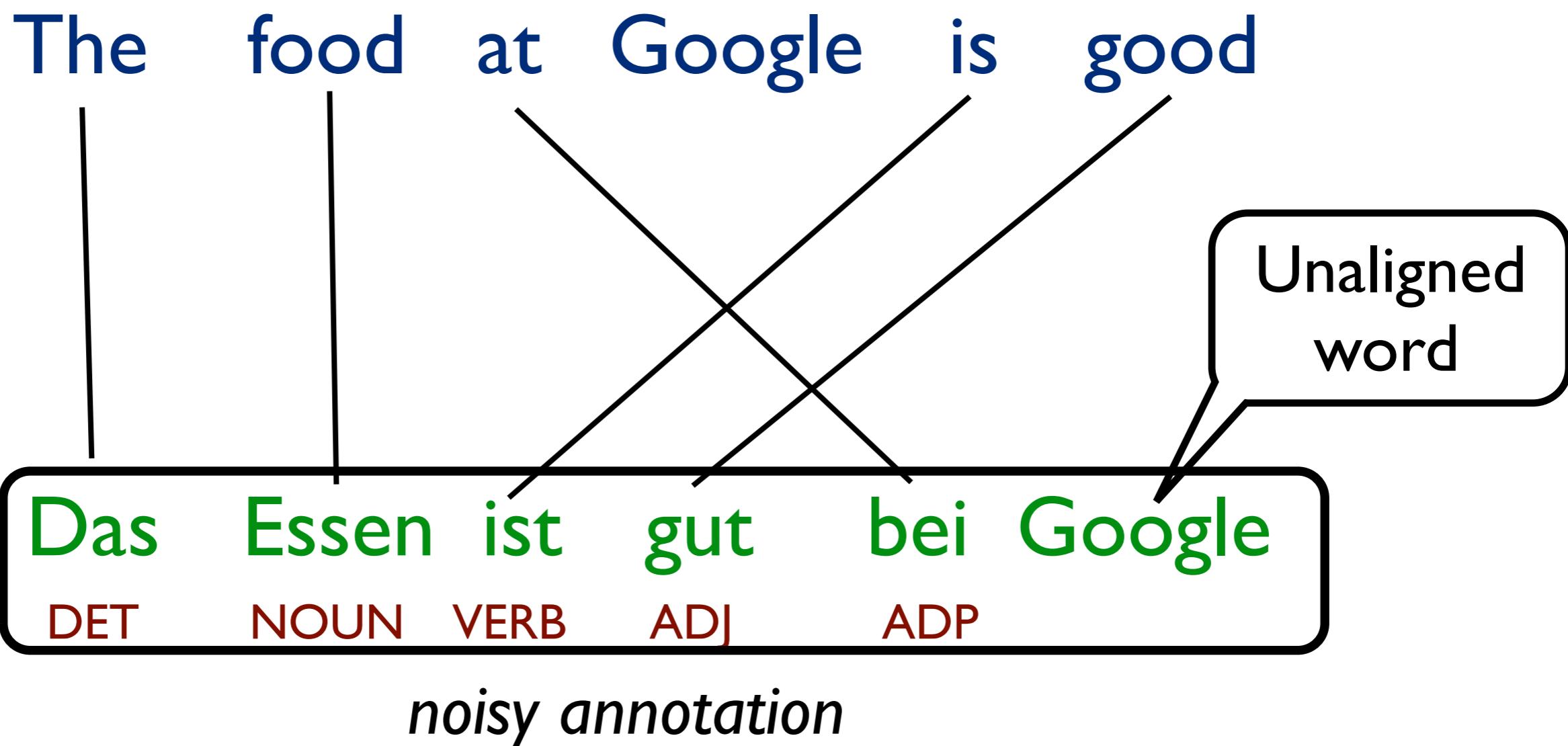


The food at Google is good

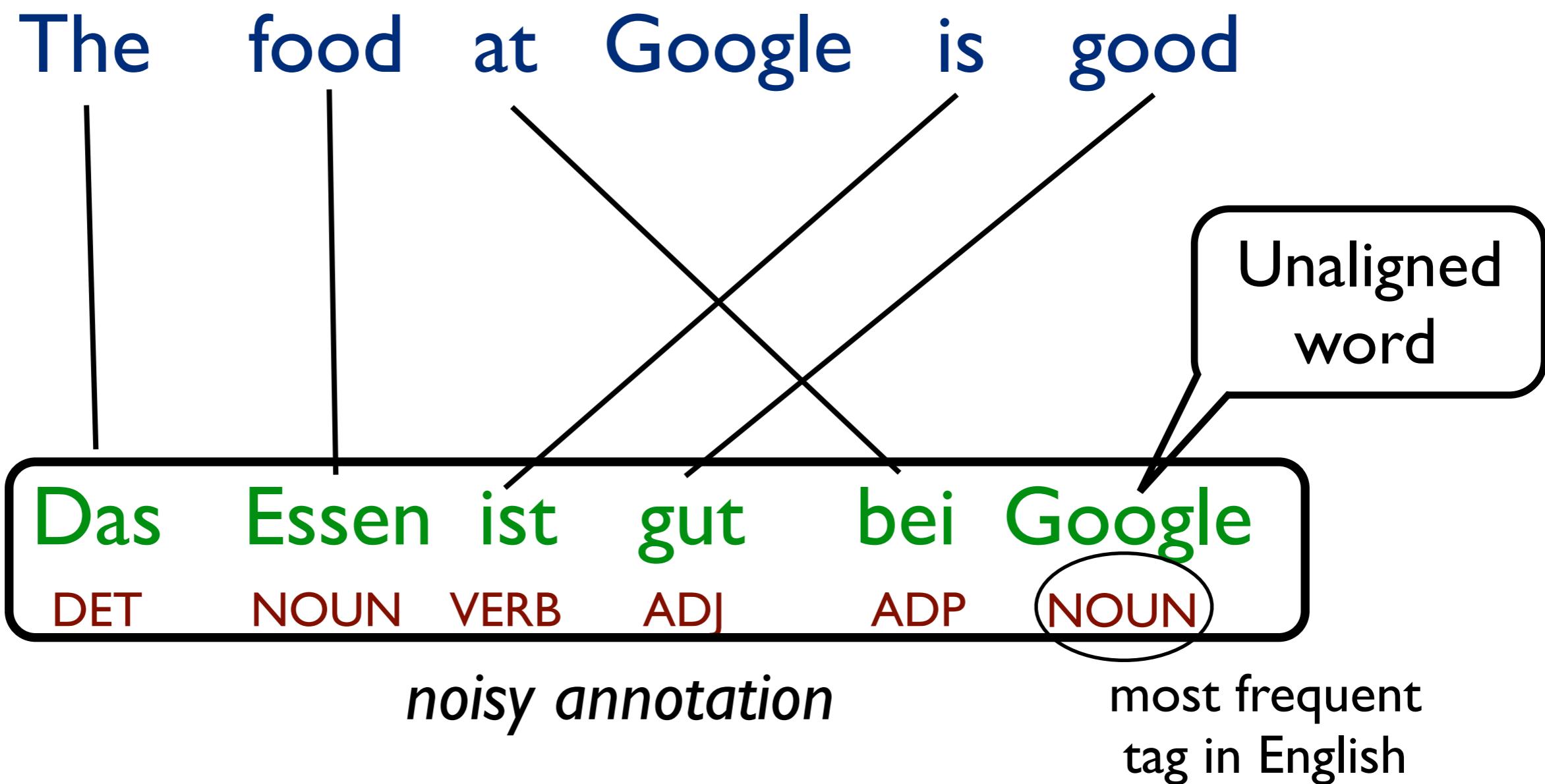


noisy annotation

Cross-Lingual Projection



Cross-Lingual Projection



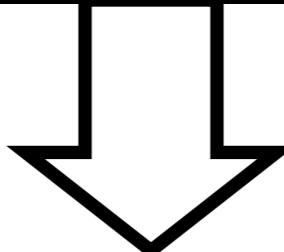
Cross-Lingual Projection



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

+

more projected tagged sentences



supervised training

direct projection **tagger**

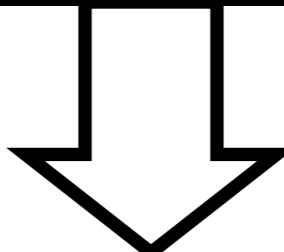
Cross-Lingual Projection



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

+

more projected tagged sentences



supervised training

direct projection **tagger**

Model I : Direct Projection

Model I : Direct Projection

									Average
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	
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)

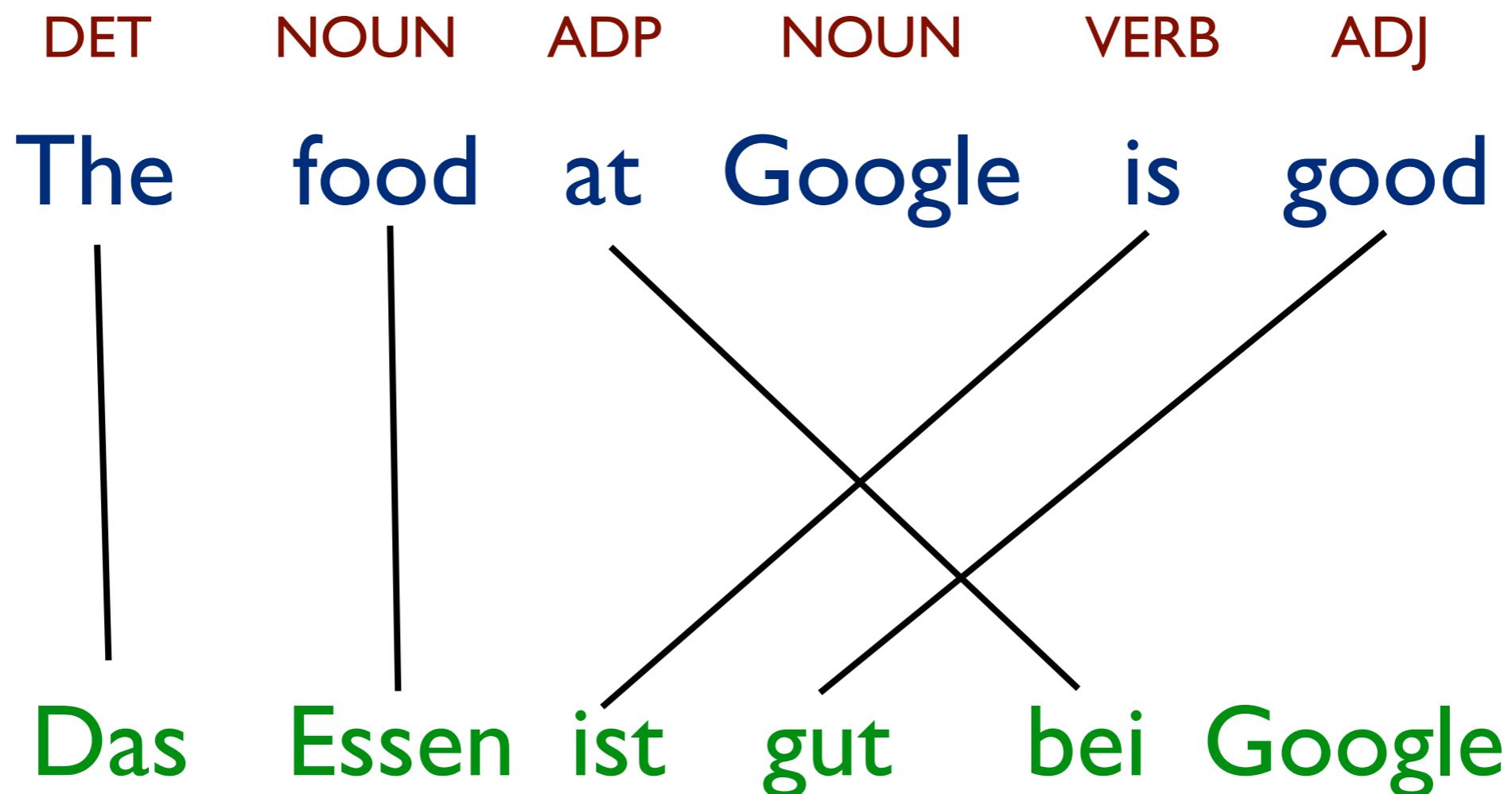
Model I : Direct Projection

									Average
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	
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

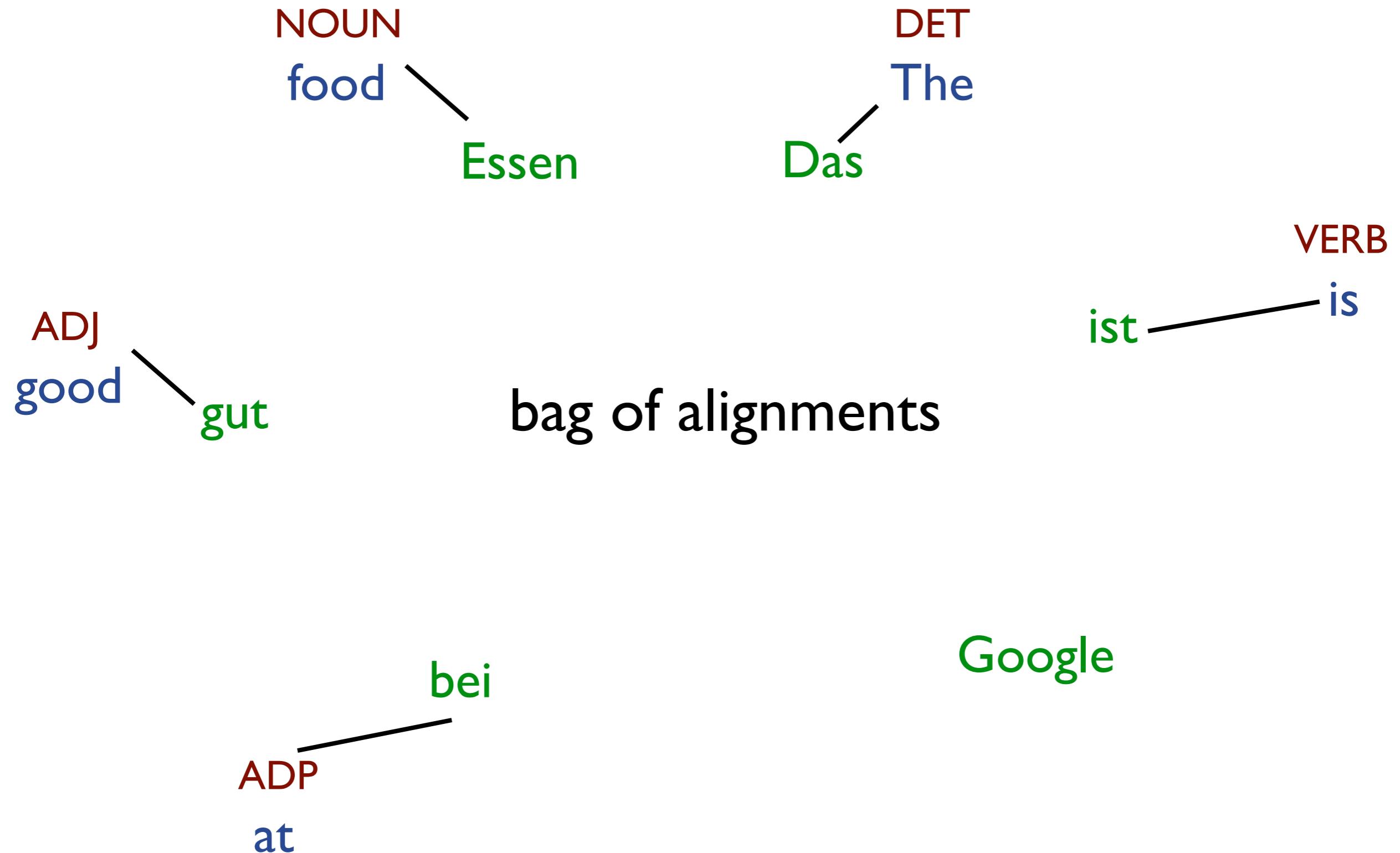
consistent improvements

Yarowsky and Ngai (2001)

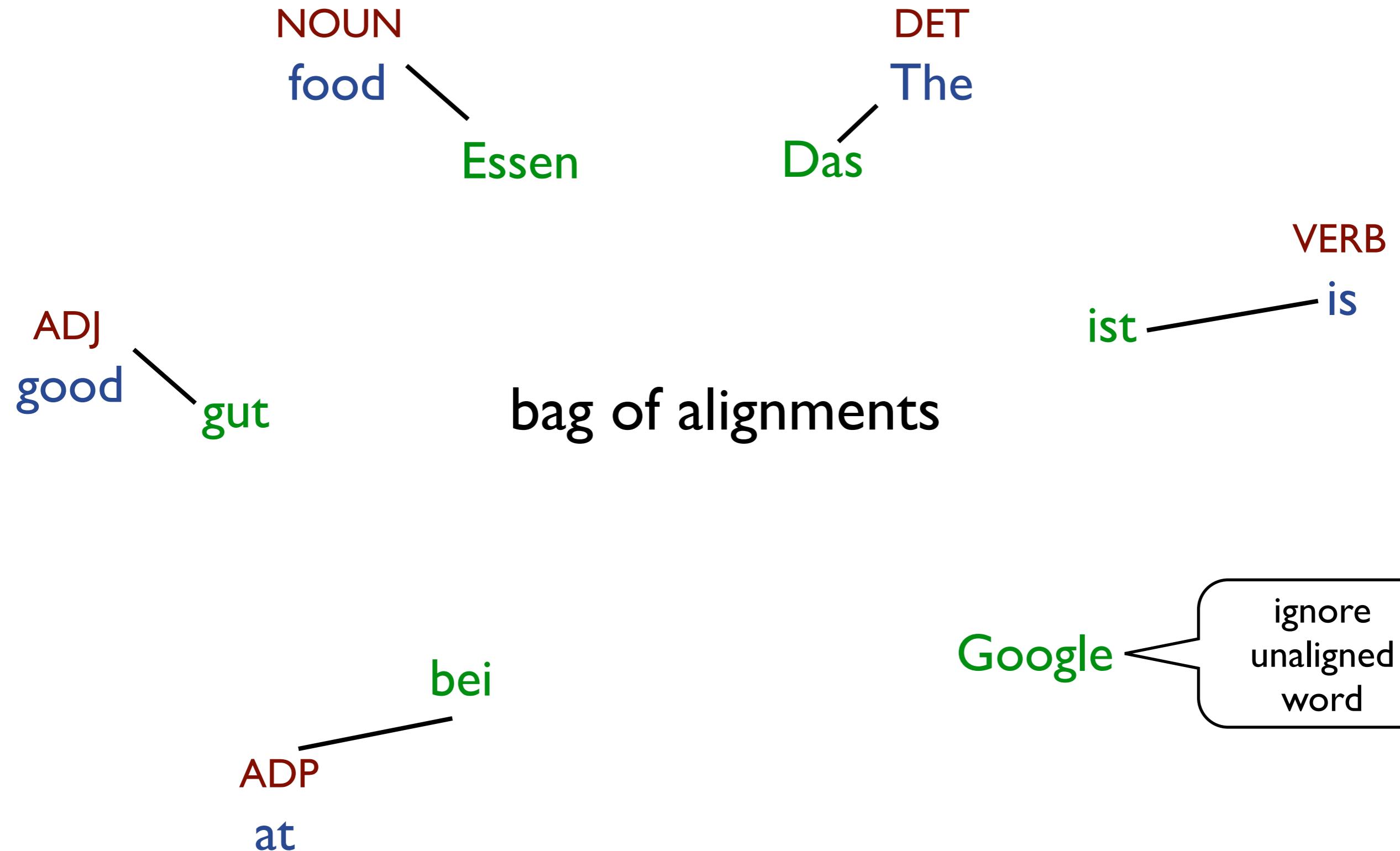
Cross-Lingual Projection



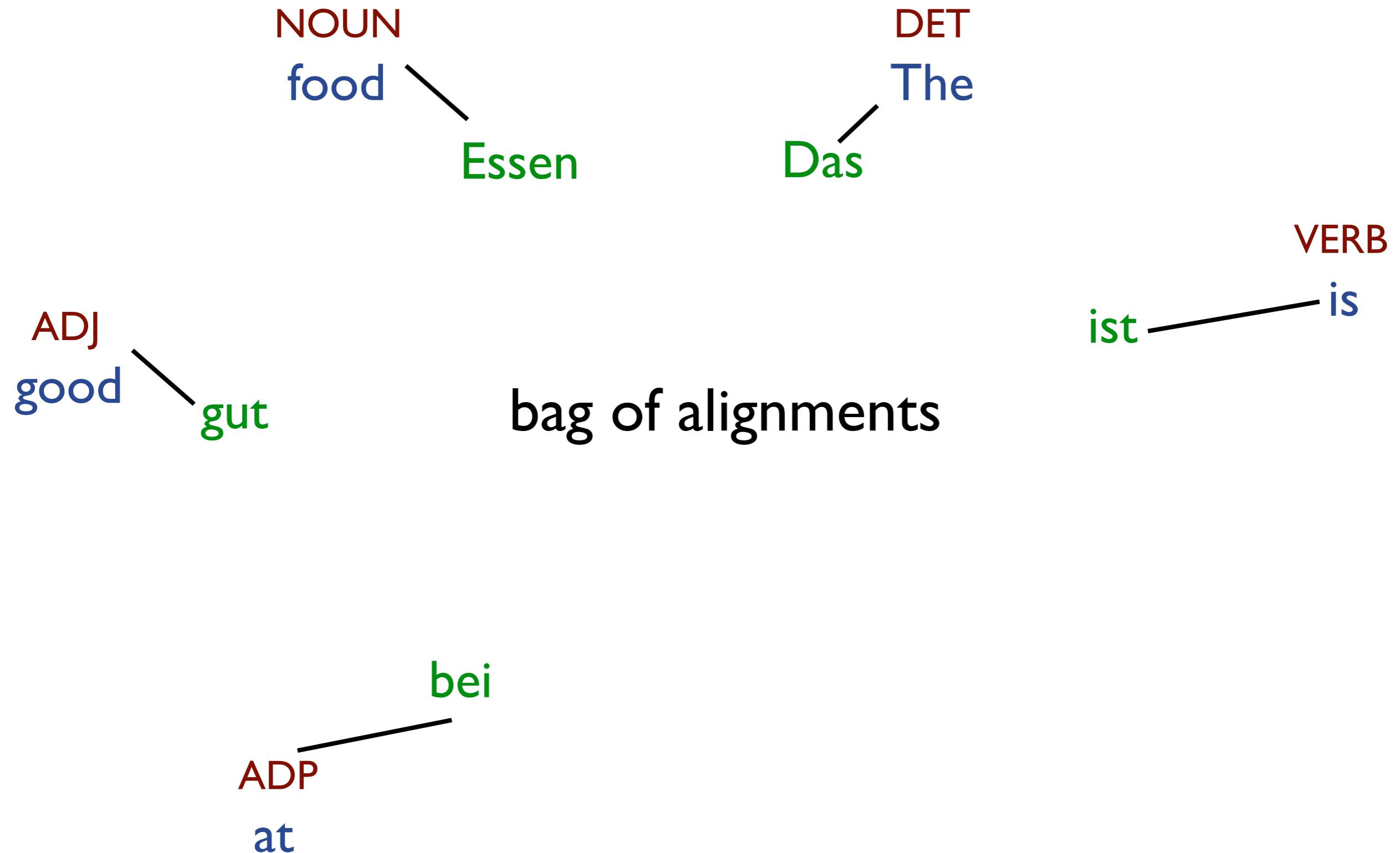
Cross-Lingual Projection



Cross-Lingual Projection

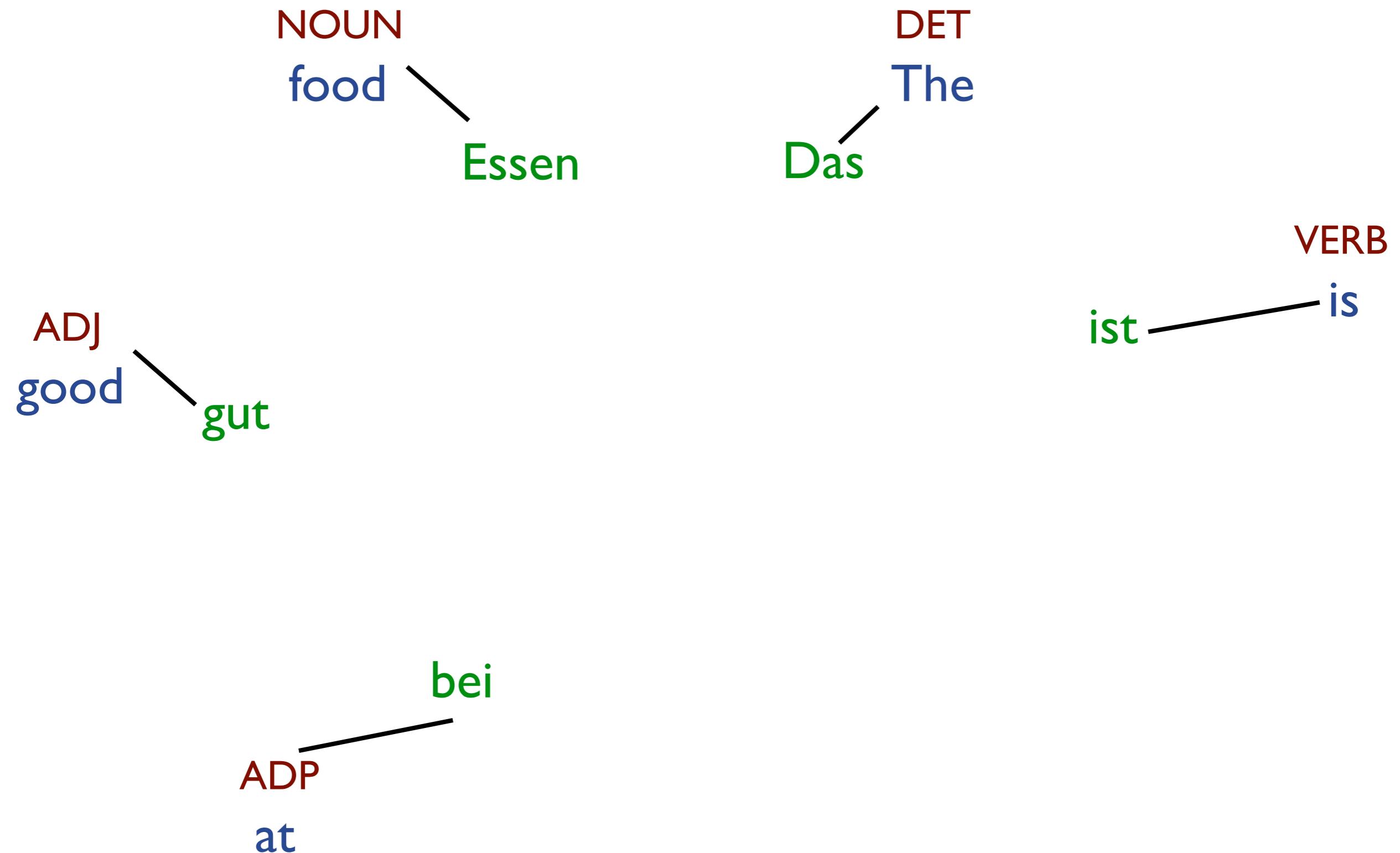


Cross-Lingual Projection

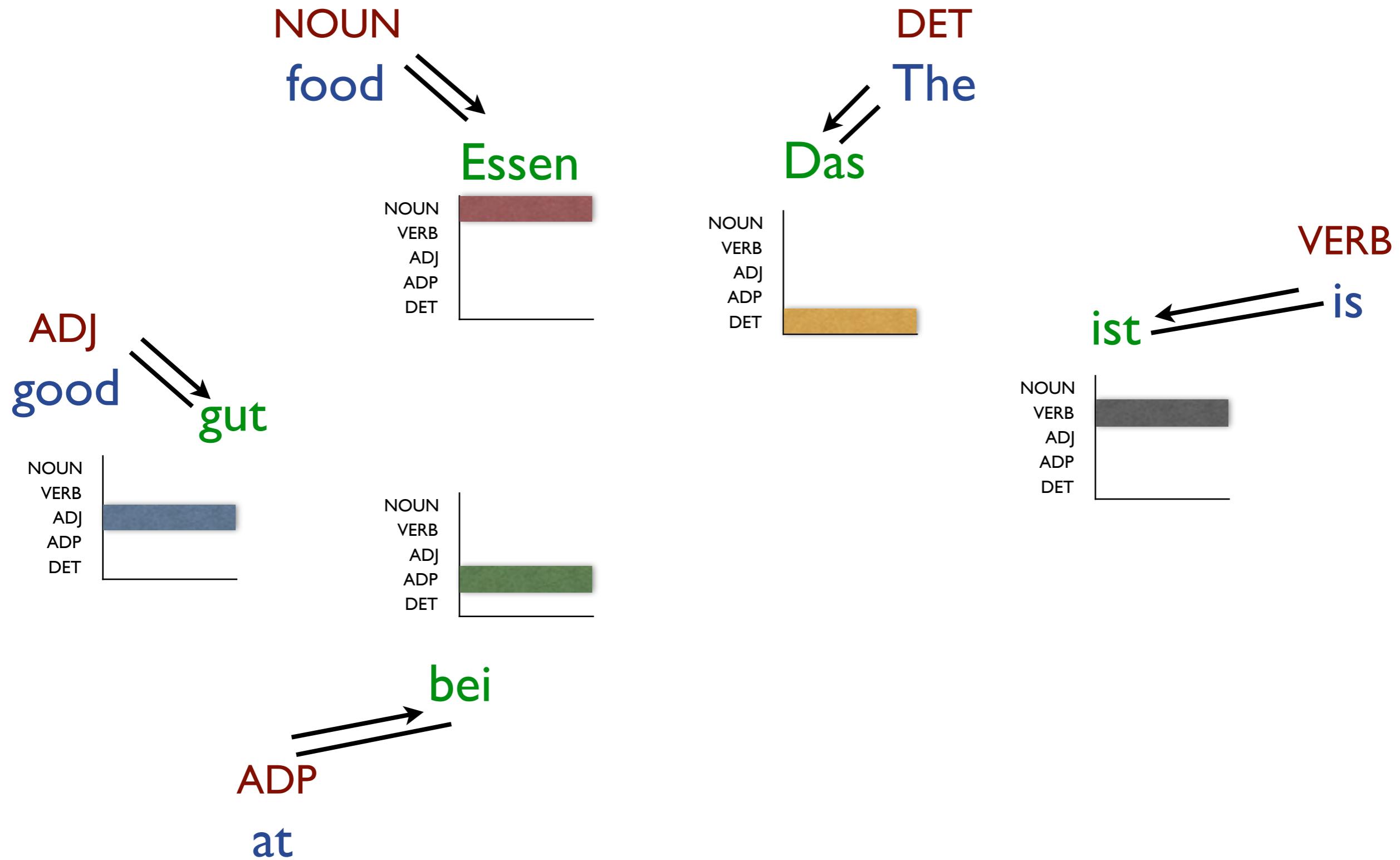


Cross-Lingual Projection

Google™

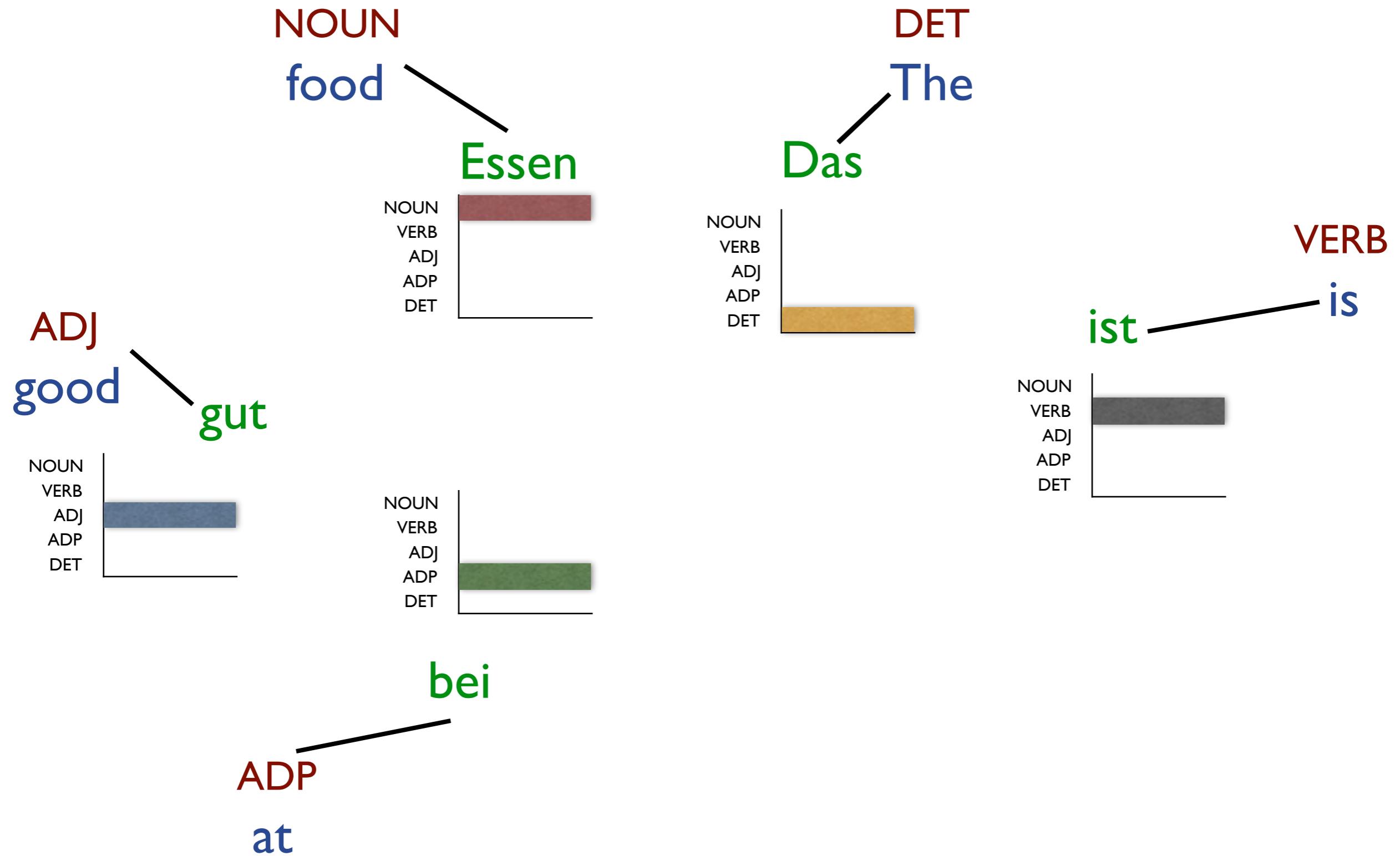


Cross-Lingual Projection



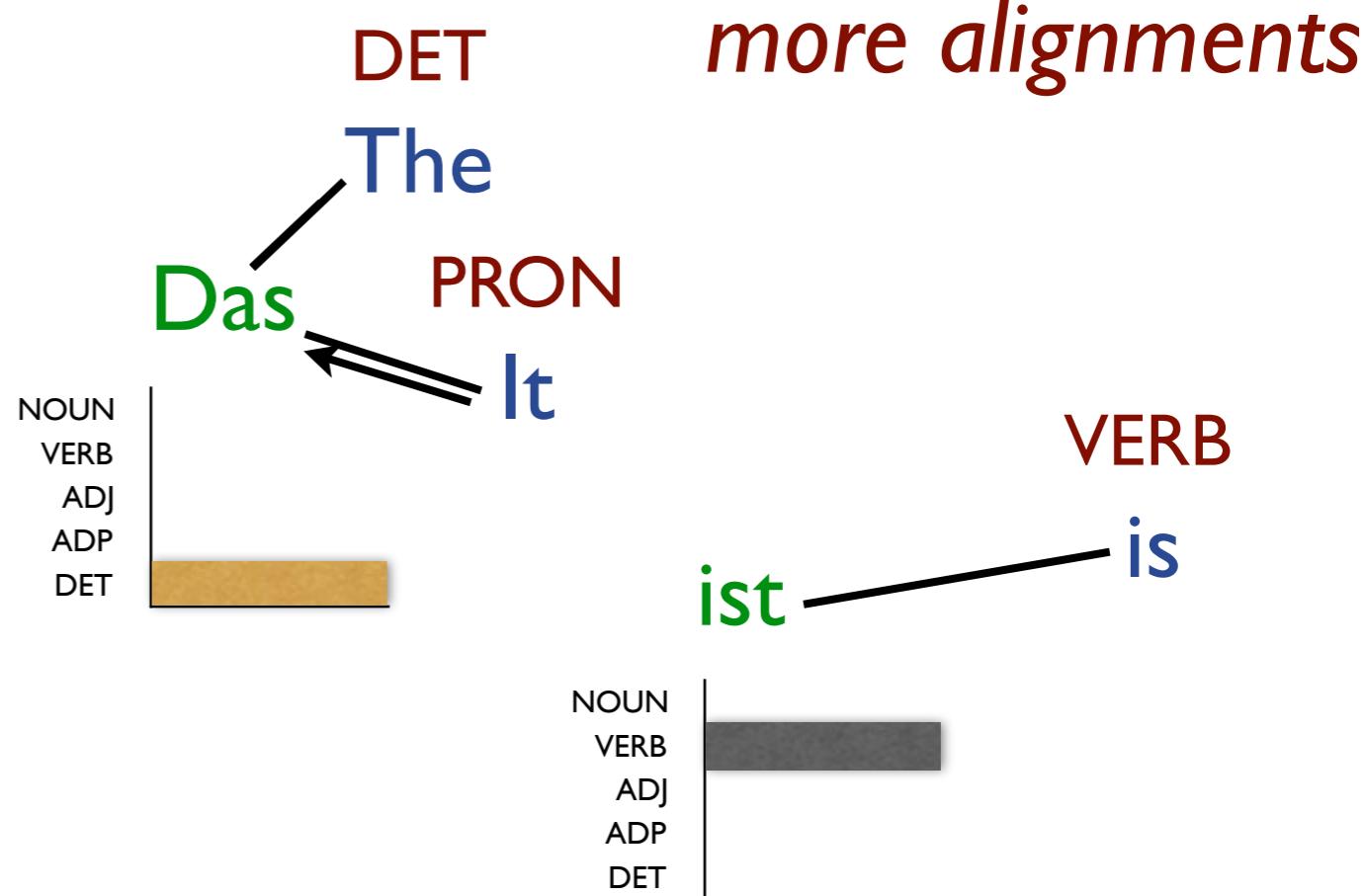
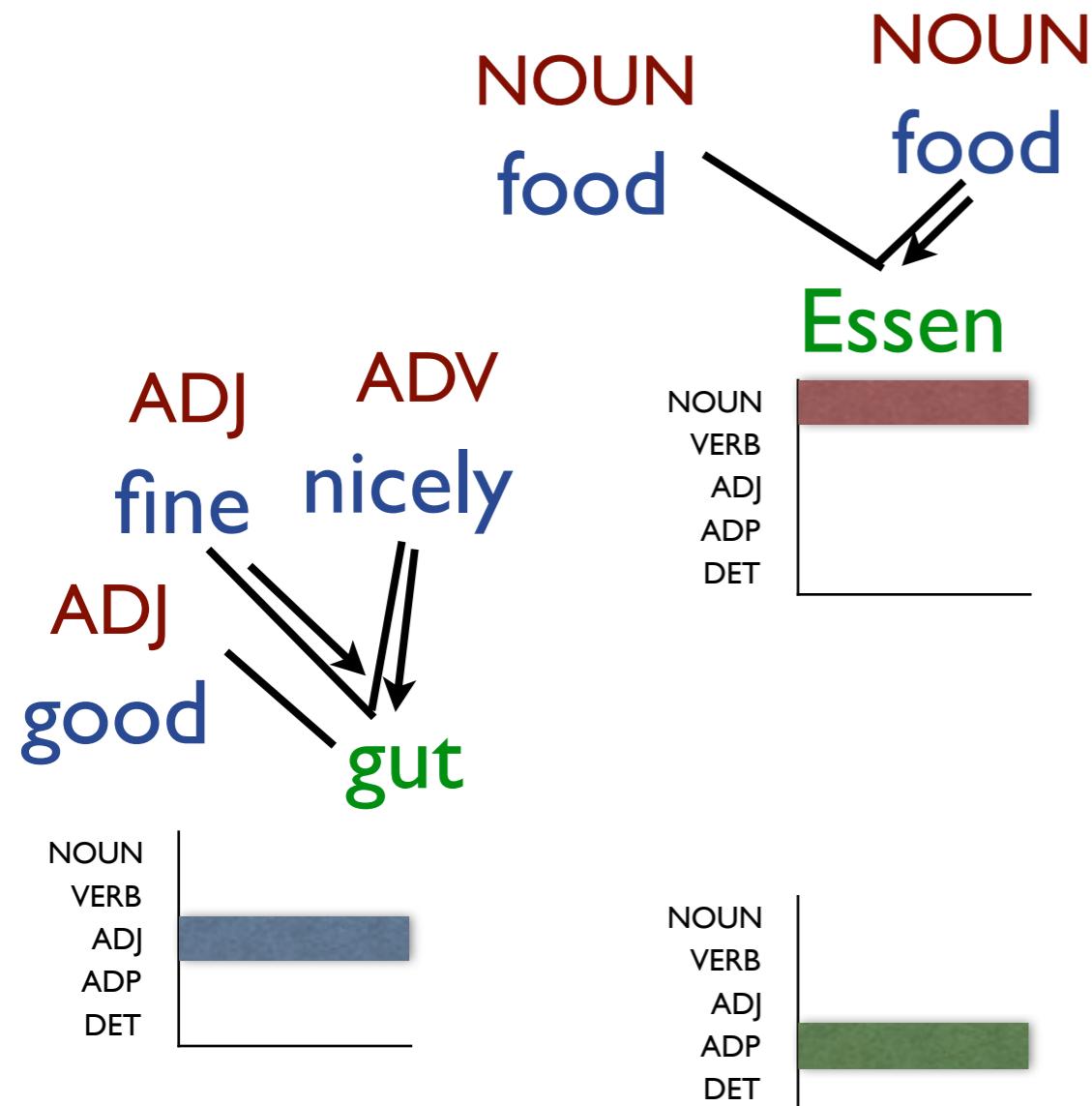
Cross-Lingual Projection

Google™



Cross-Lingual Projection

Google™



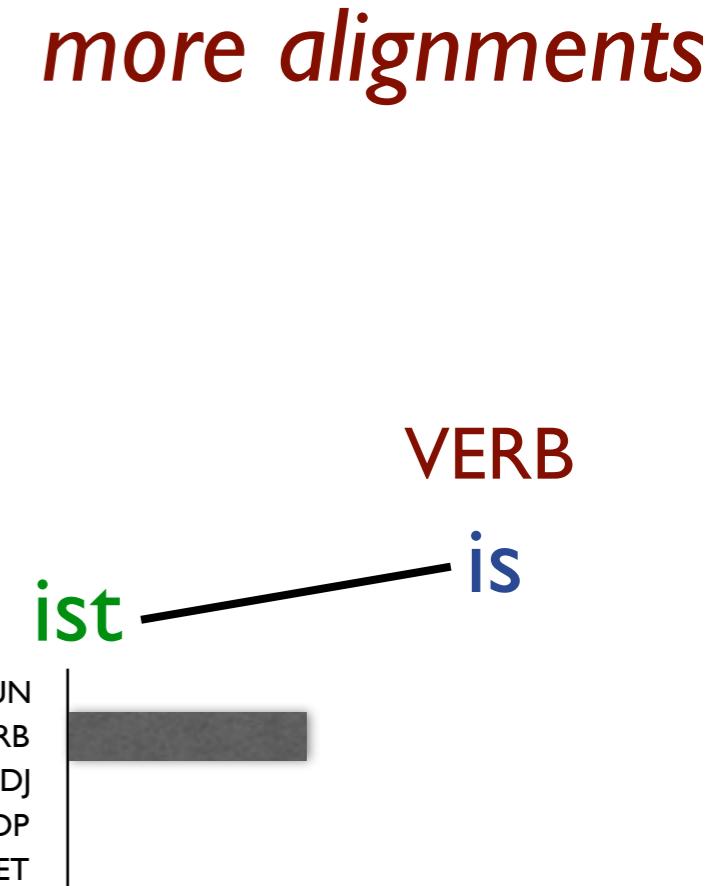
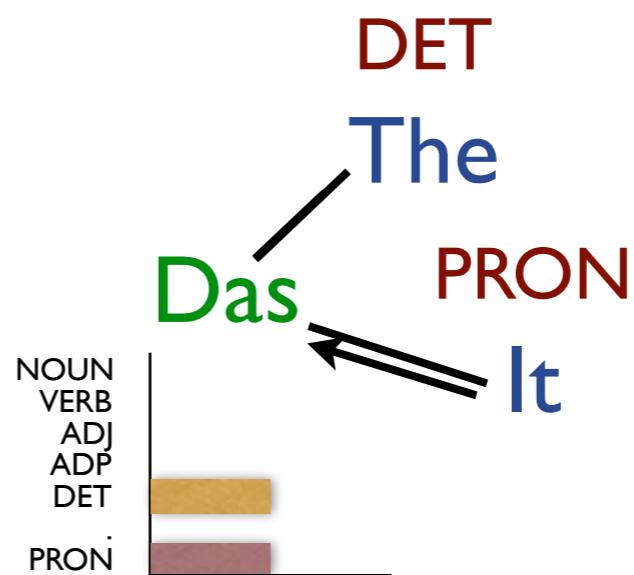
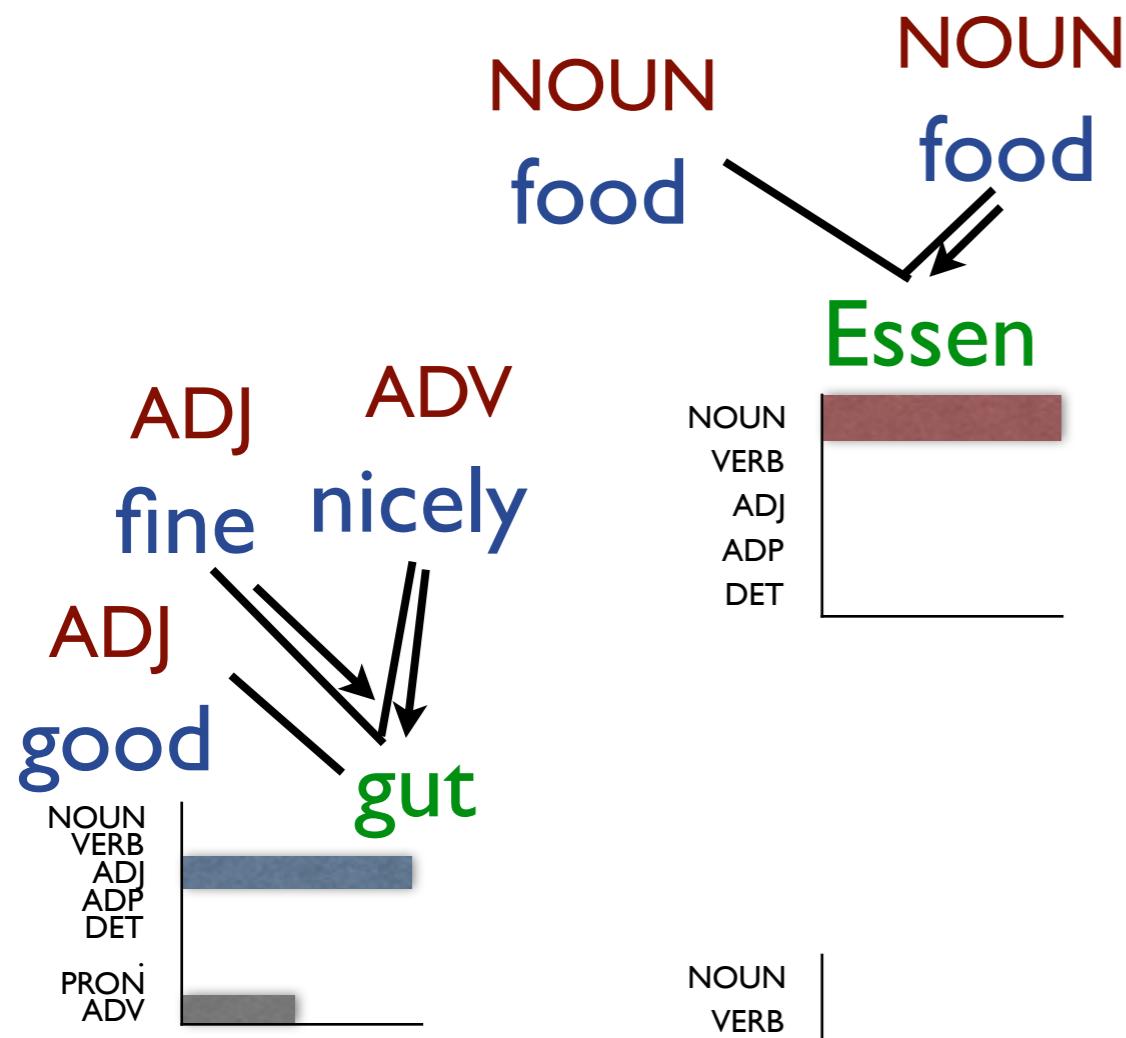
bei
ADP
at

more alignments

VERB
is

Cross-Lingual Projection

The Google logo is displayed in its signature multi-colored, rounded font. The letters are arranged as follows: G (blue), o (red), o (yellow), g (green), l (blue), e (red). A small trademark symbol (TM) is located at the top right of the 'e'.

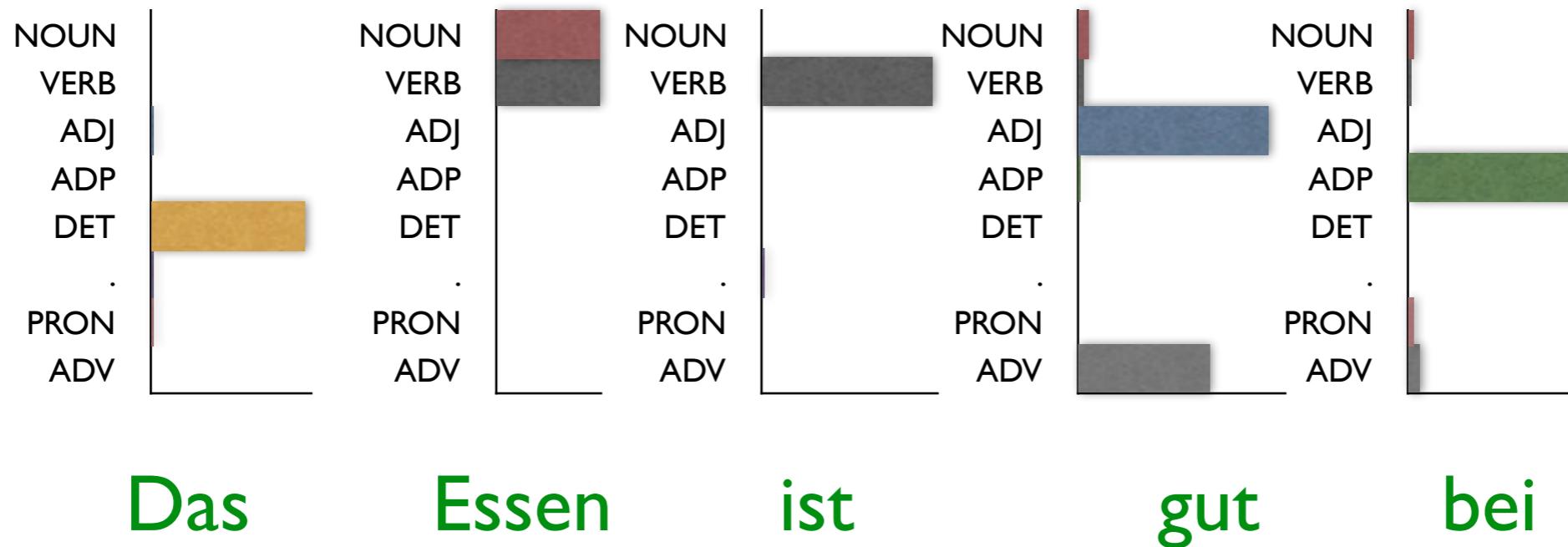


ADP
at

Cross-Lingual Projection



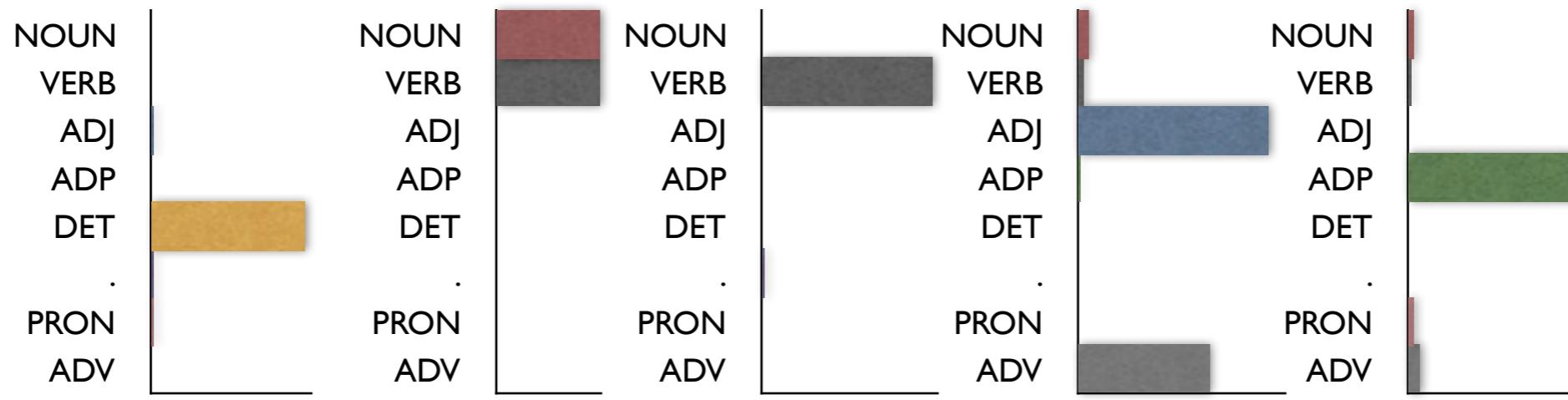
After scanning all the parallel data:



Cross-Lingual Projection



After scanning all the parallel data:



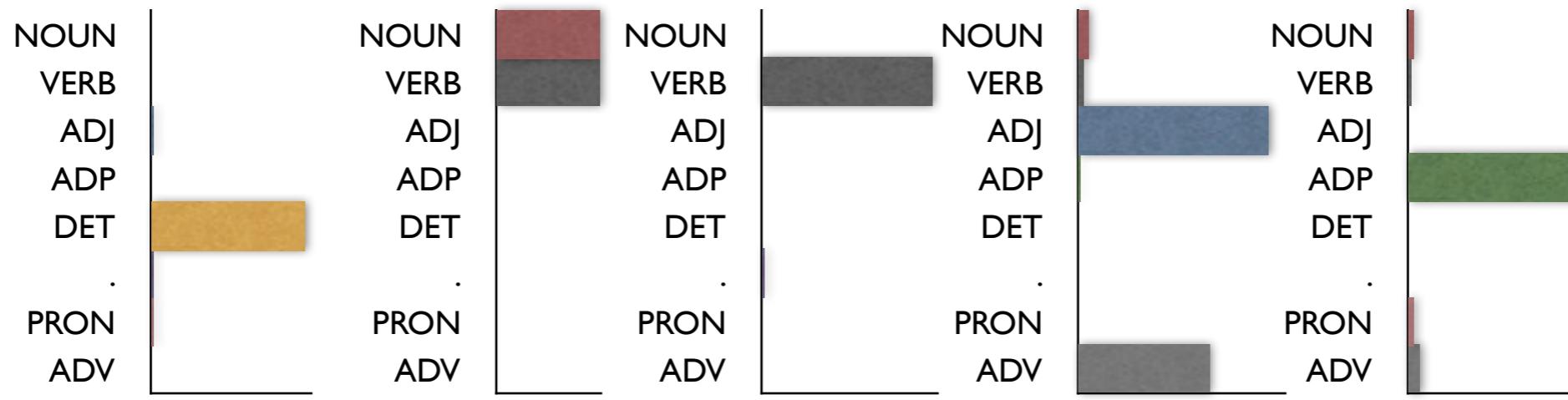
Das Essen ist gut bei

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

Cross-Lingual Projection

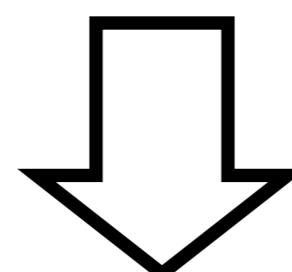


After scanning all the parallel data:



Das Essen ist gut bei

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

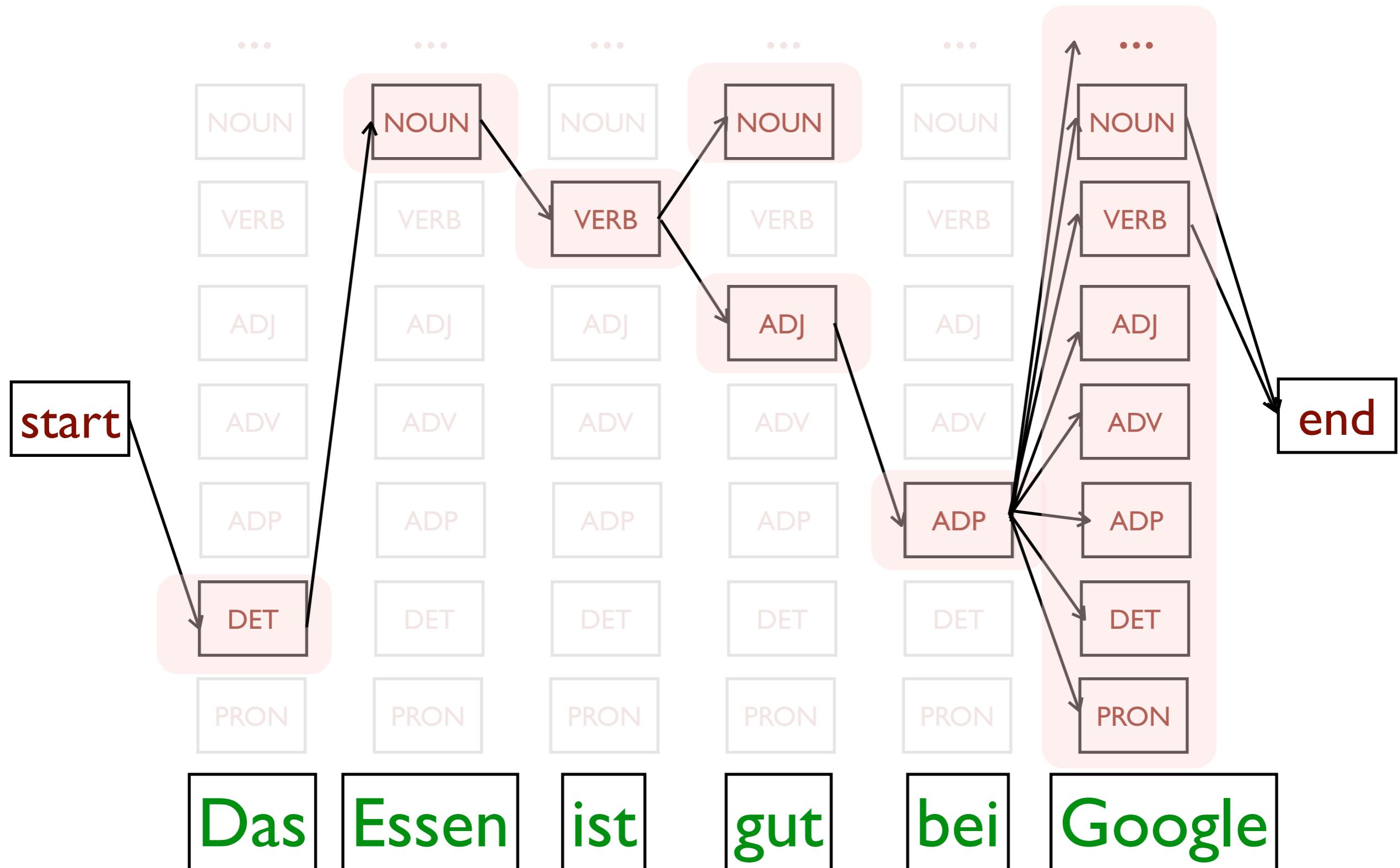


$$\text{Dict}(x) = \{y : p(y|x) > \tau\}$$

Cross-Lingual Projection



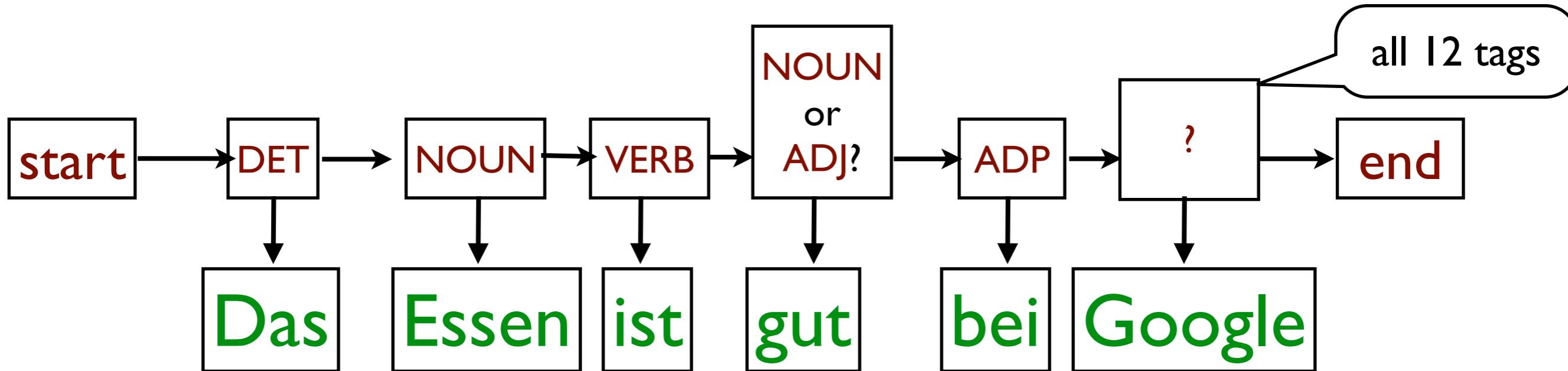
HMM with *locally-normalized log-linear models* + projected lexicon constraints



Unsupervised POS Tagging with **Projected Dictionaries**



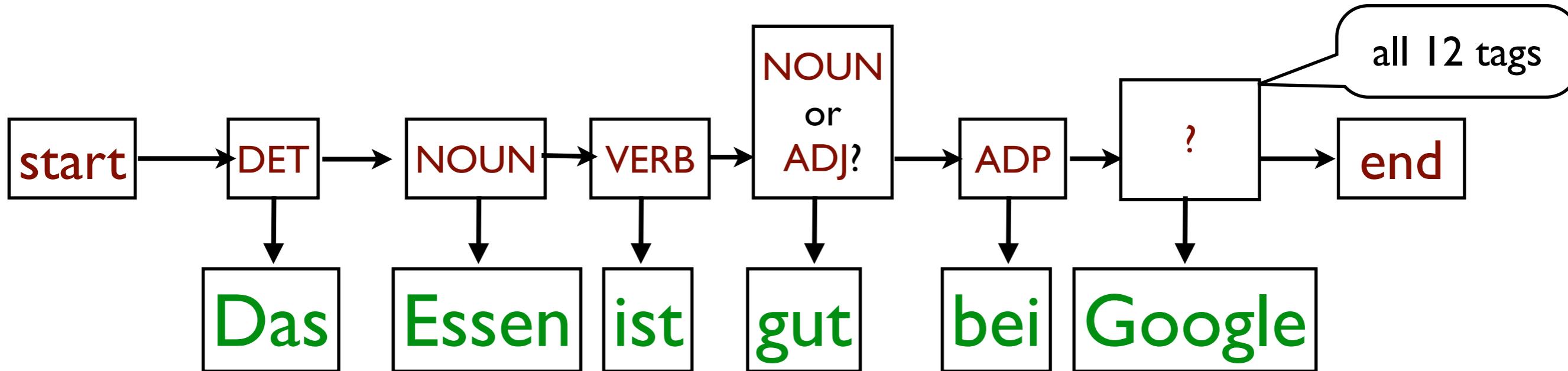
HMM with *locally-normalized log-linear models* + projected lexicon constraints



Unsupervised POS Tagging with **Projected Dictionaries**



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

									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

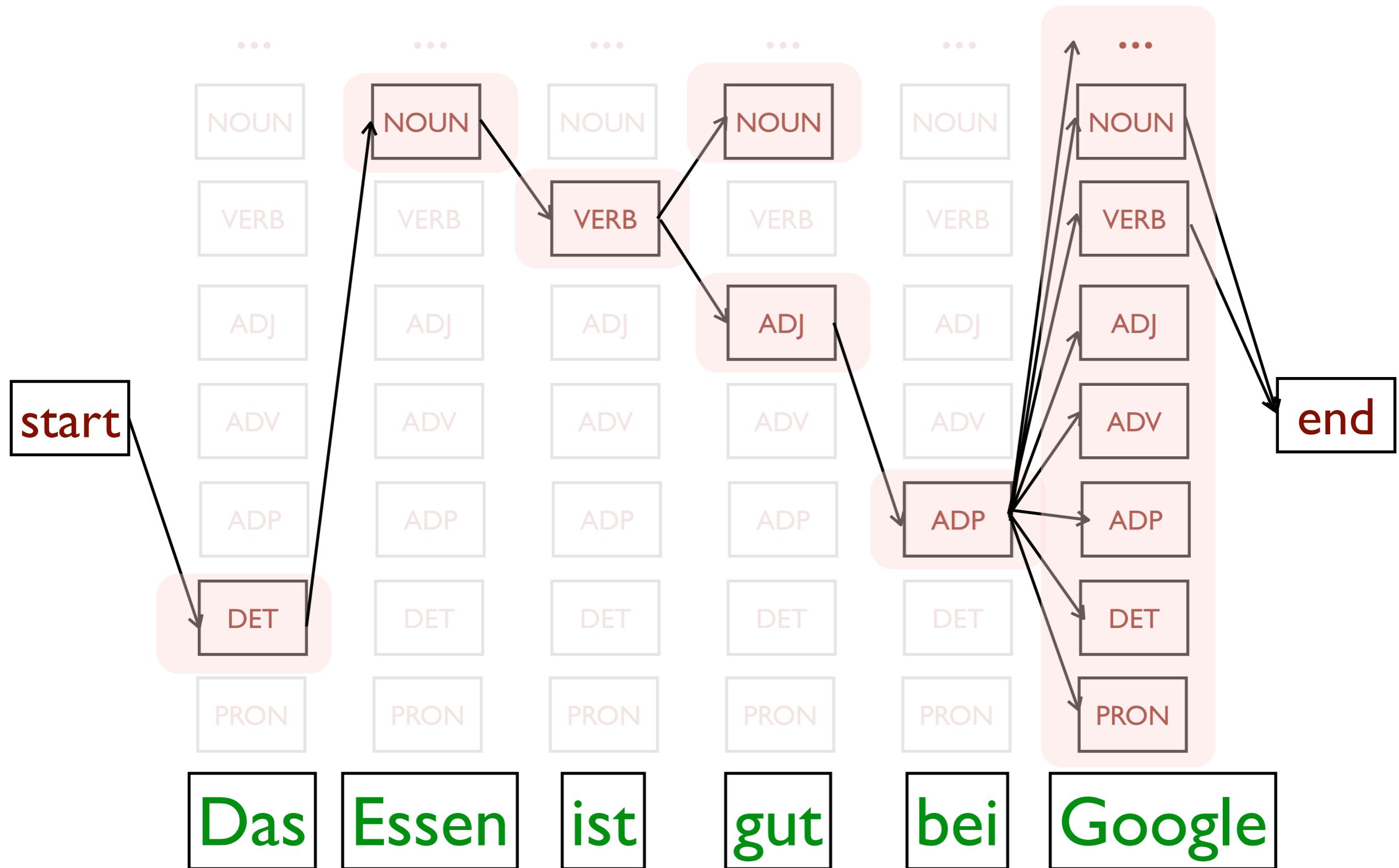
									Average
EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
Projected Dictionary	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3

Improvement over direct projection for most languages

Unsupervised POS Tagging with **Projected Dictionaries**



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

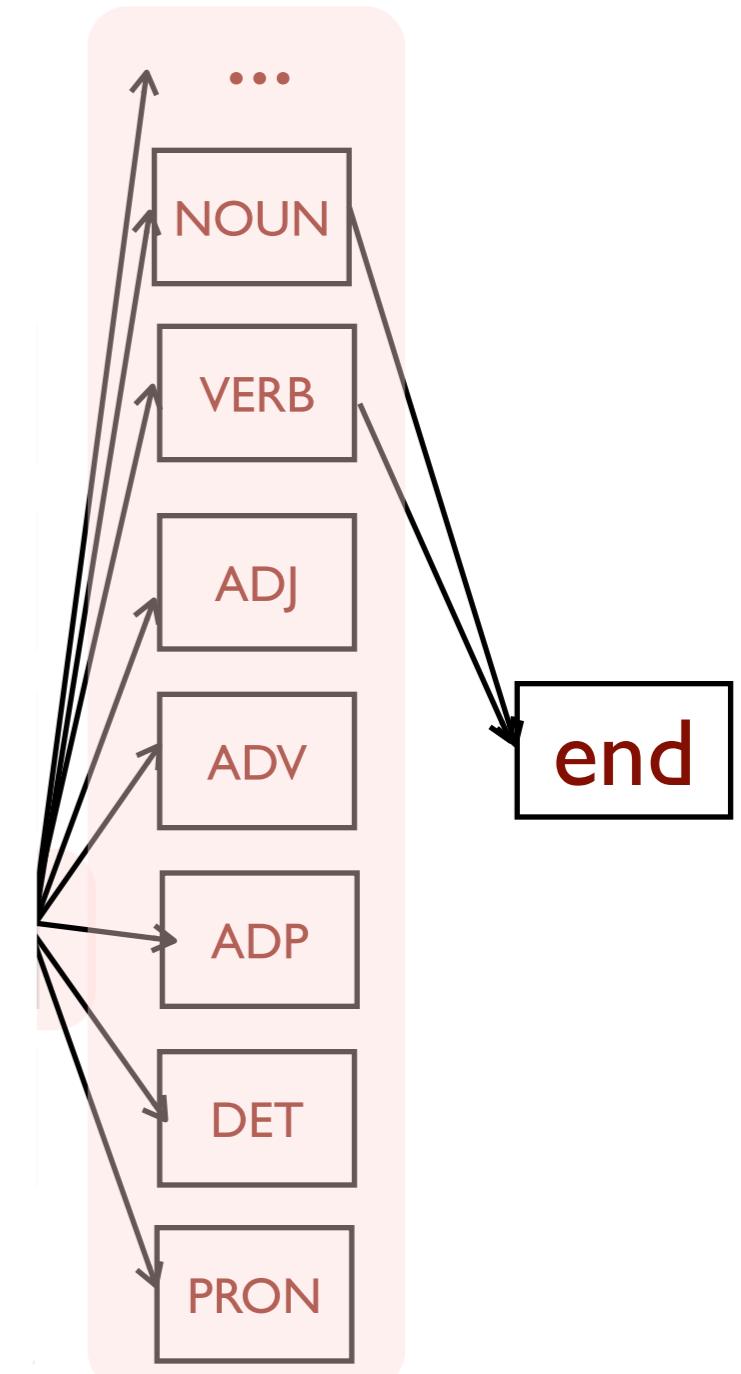


HMM with *locally-normalized log-linear models* + projected lexicon constraints

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

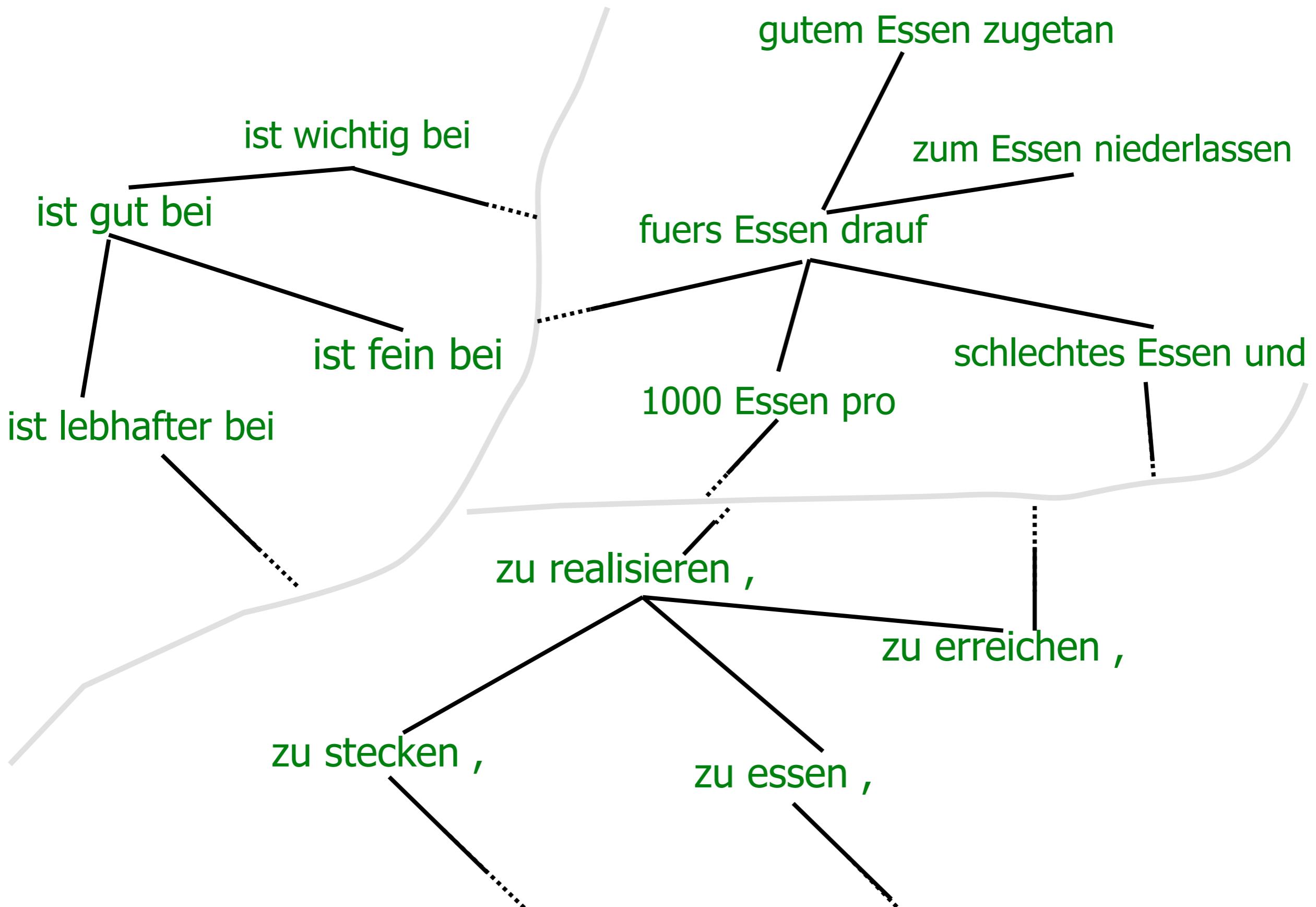
Graph-Based Semi-Supervised Learning



- 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

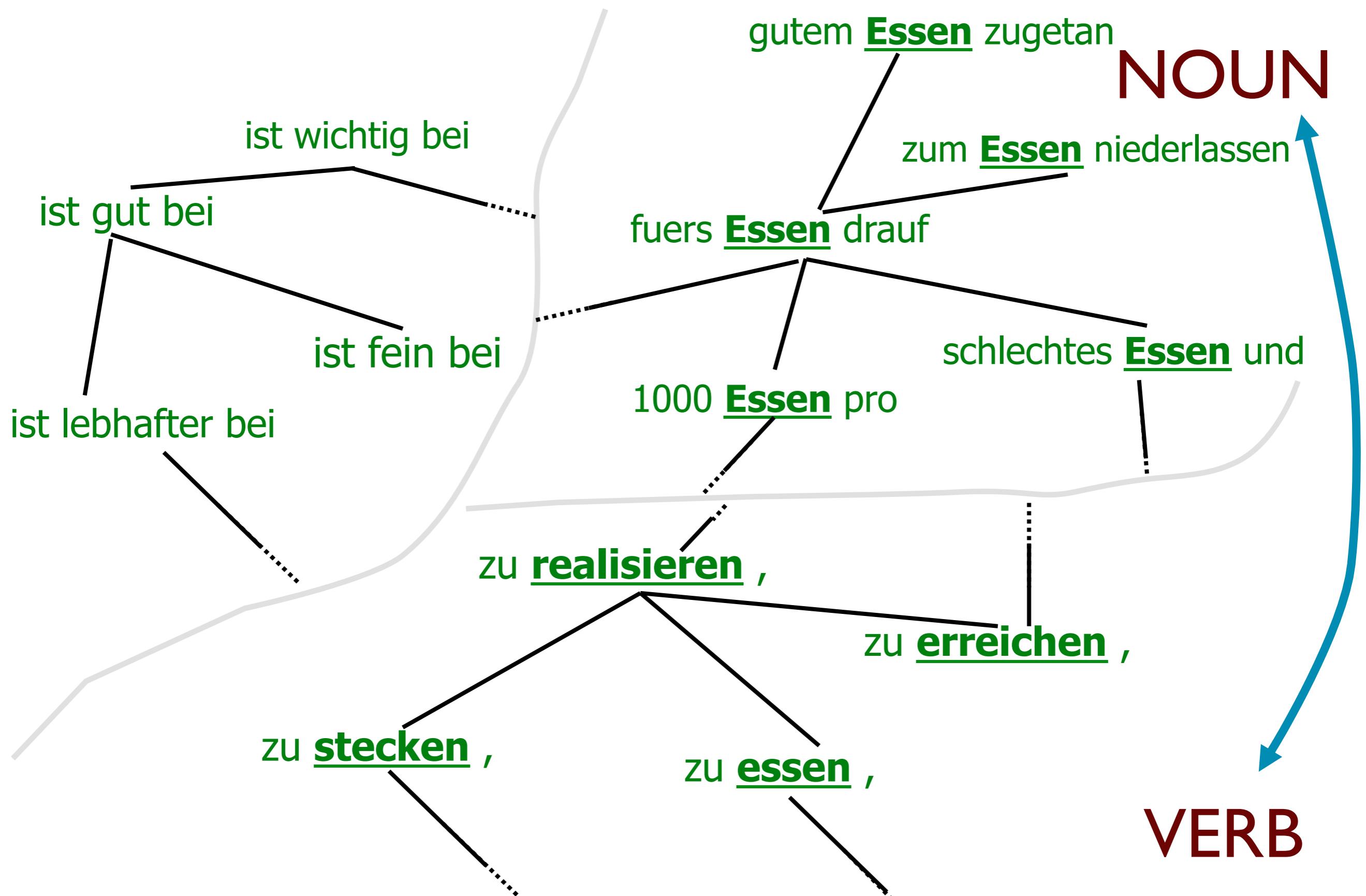
Example Graph in German

Google™



Example Graph in German

Google™



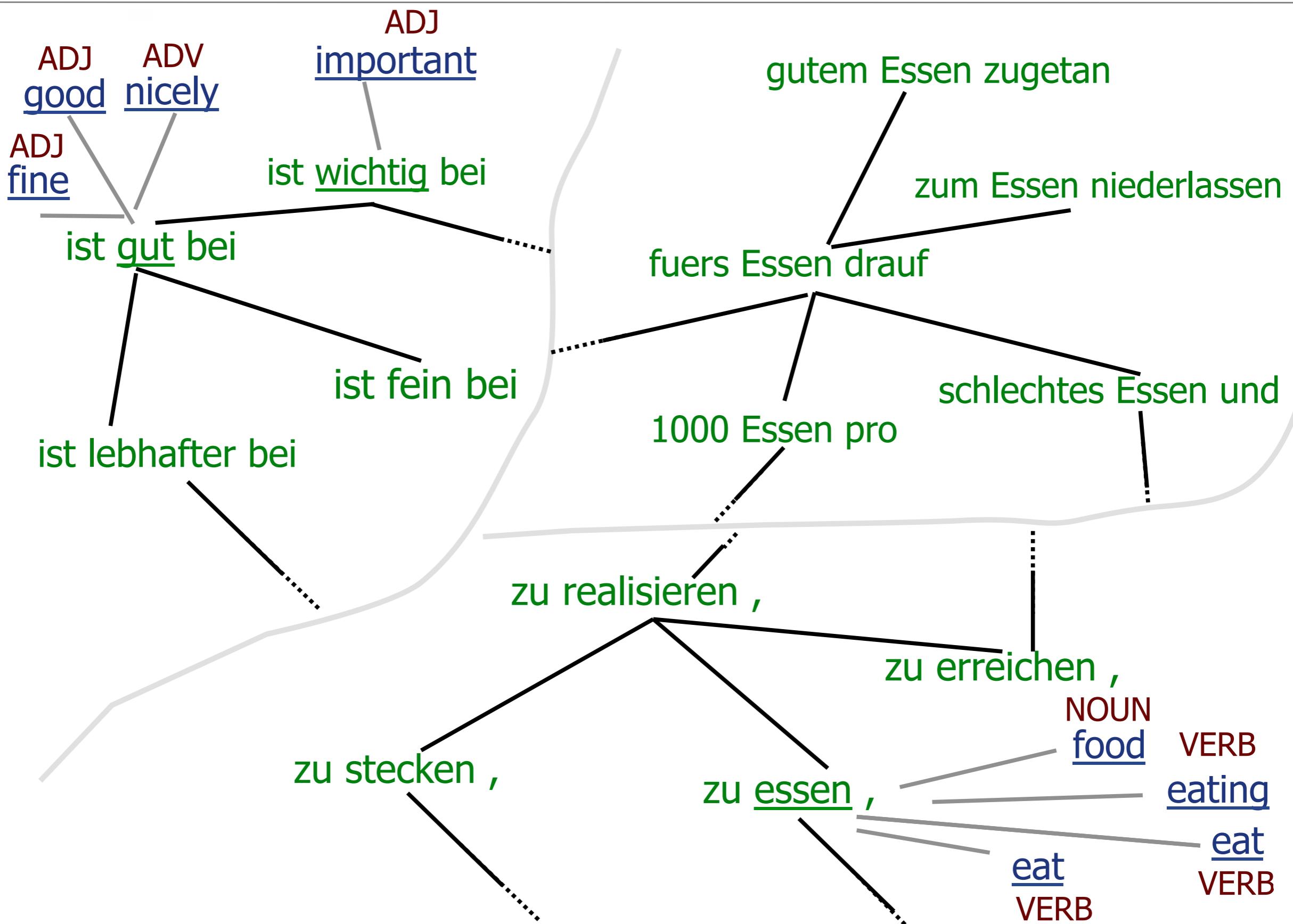
Graph-Based Semi-Supervised Learning



- 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

Google™

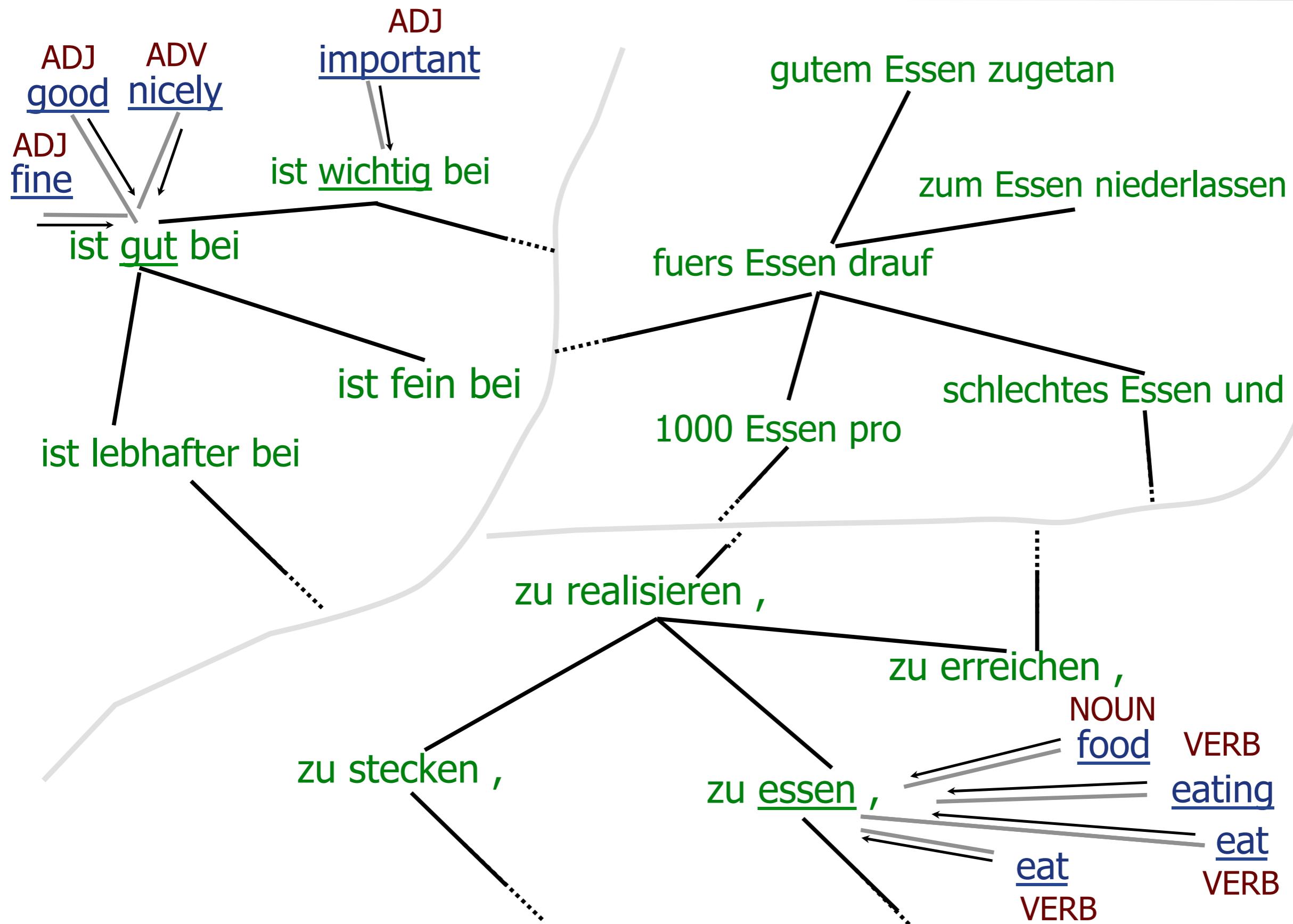


Graph-Based Semi-Supervised Learning

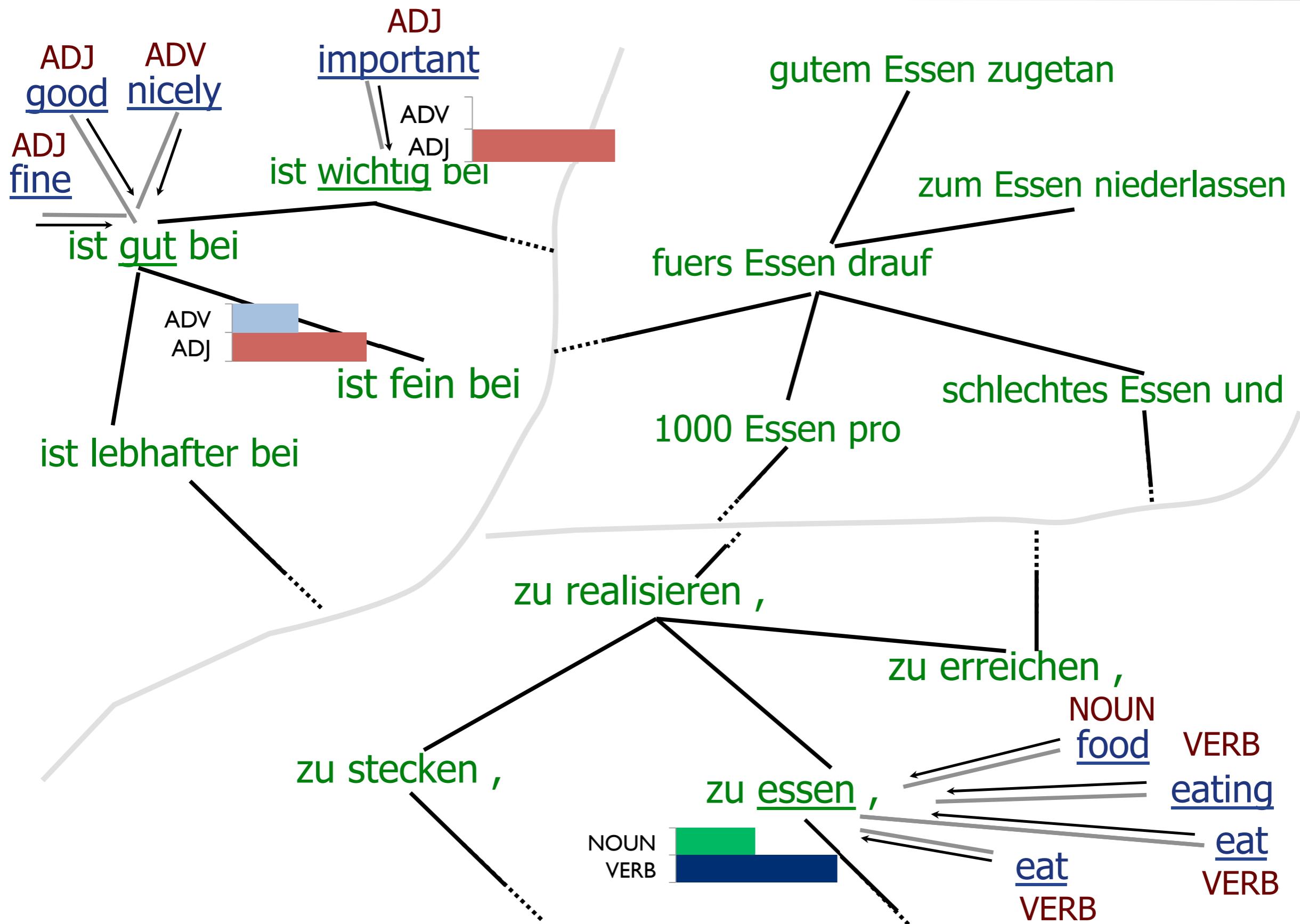


- Plug in auto-tagged words from a **source** language
- Links between source and target vertices are word alignments
- Run first stage of *label propagation*
 - Source language → Target language

Bilingual Graph-Based Label Propagation



Bilingual Graph-Based Label Propagation

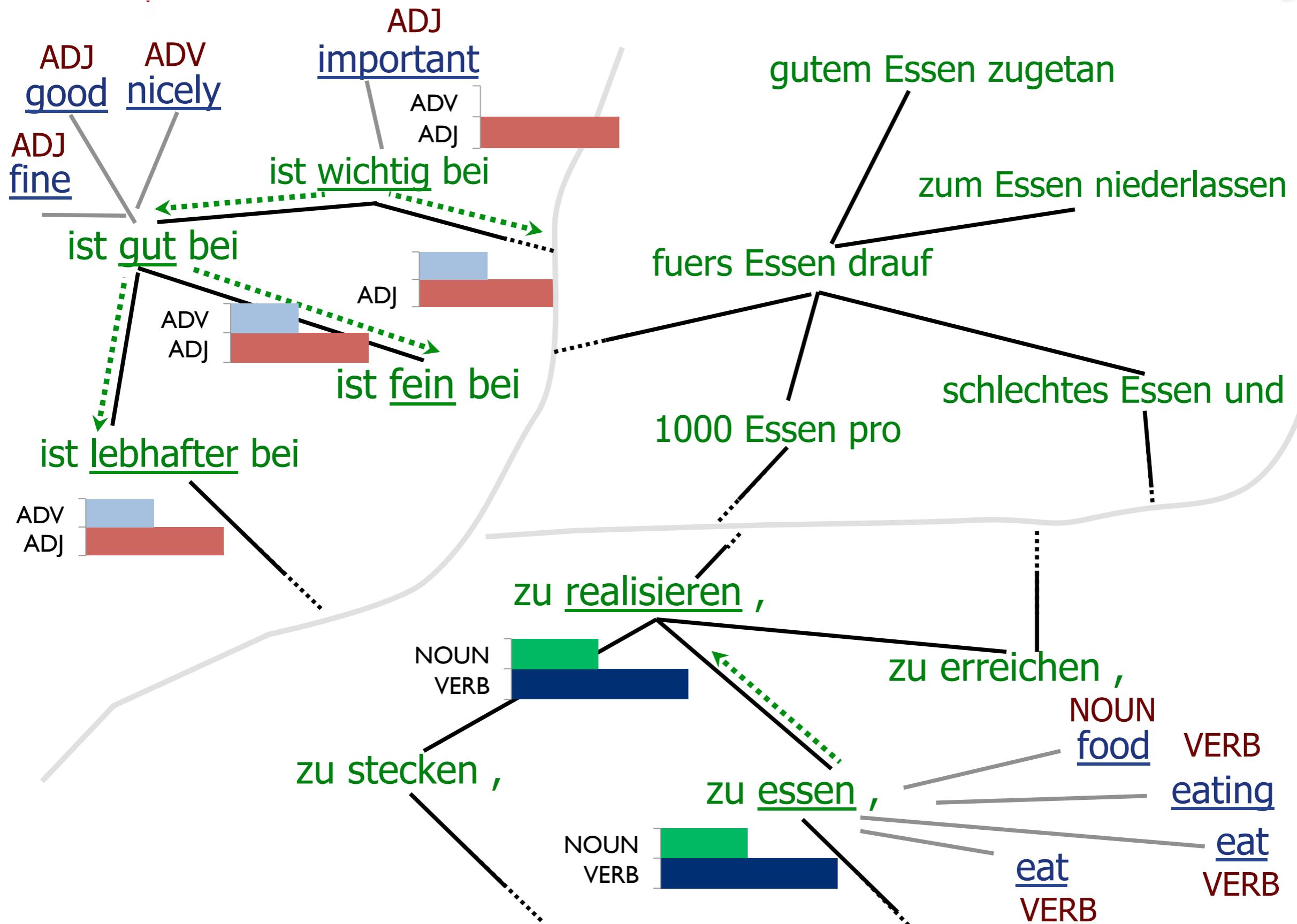


Graph-Based Semi-Supervised Learning

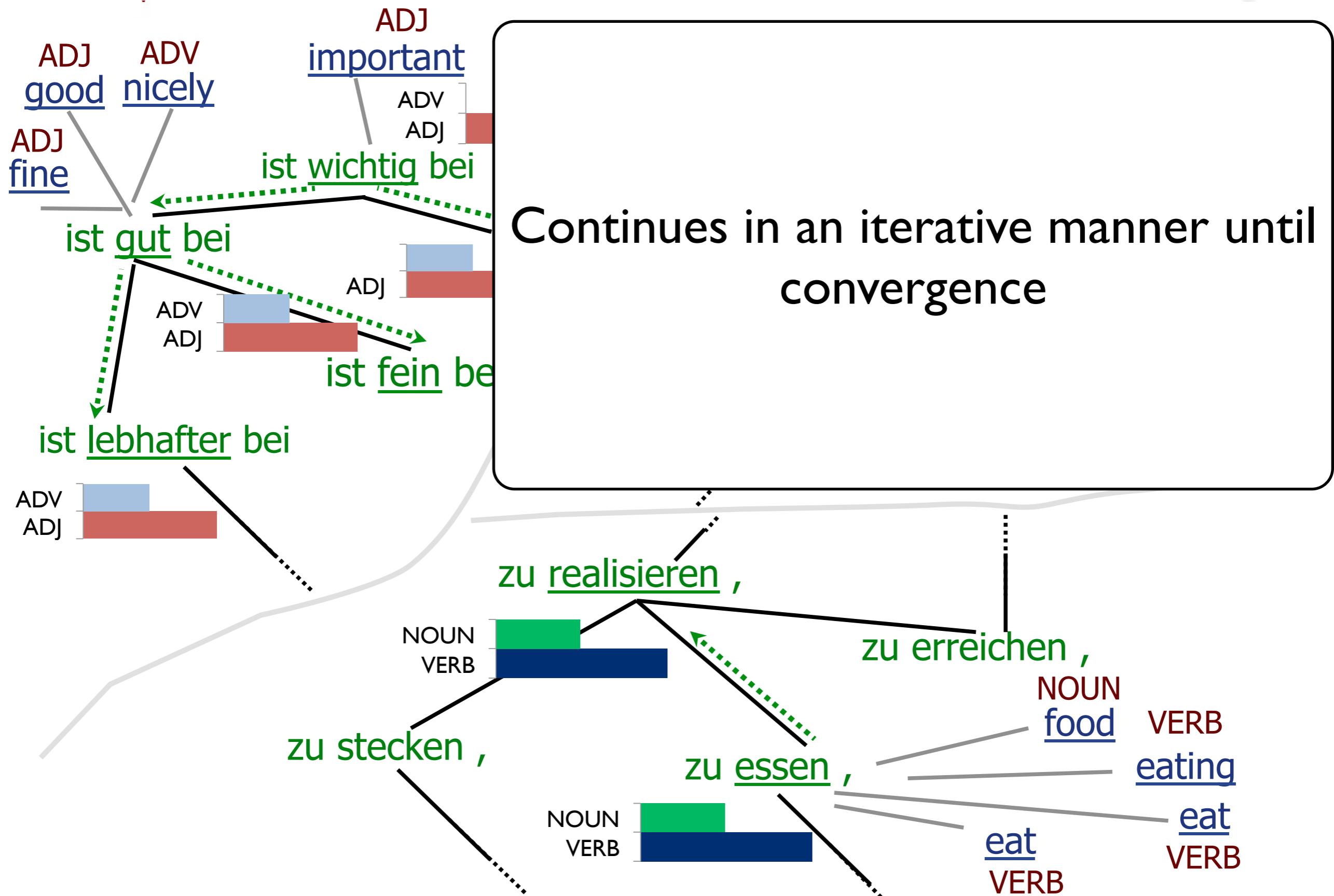


- Plug in auto-tagged words from a **source** language
- Links between source and target vertices are word alignments
- Run first stage of *label propagation*
 - Source language → Target language
- Run second stage of label propagation
 - Within target language vertices
 - Graph objective function with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

Bilingual Graph Label Propagation



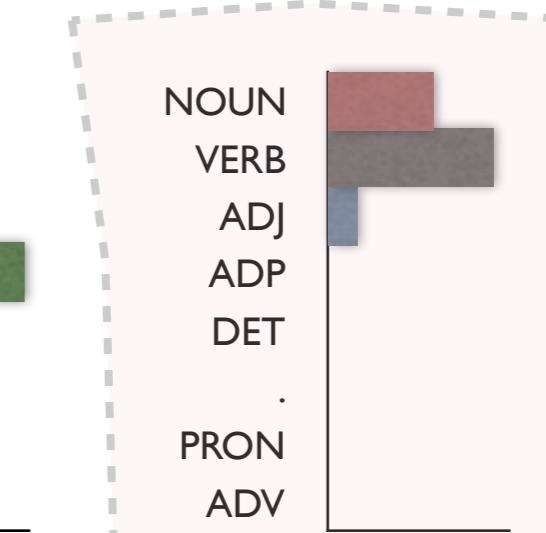
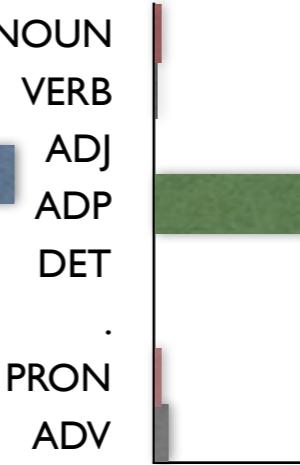
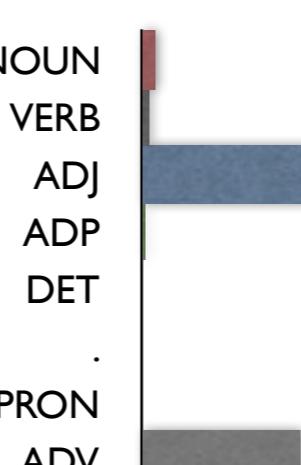
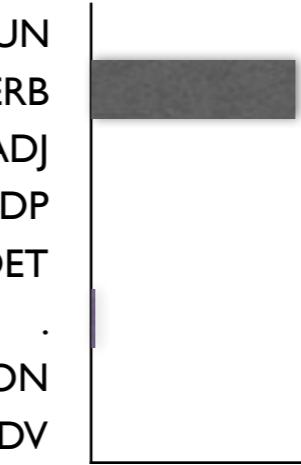
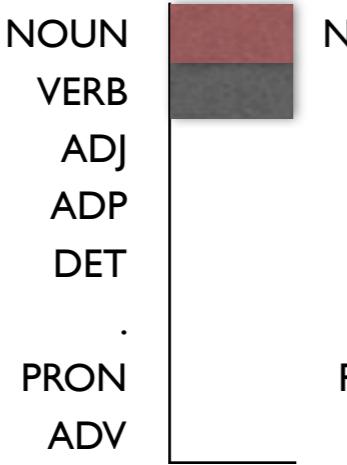
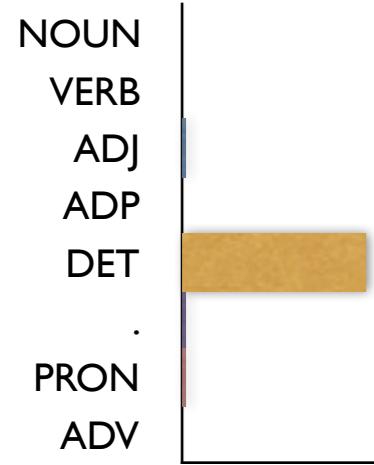
Bilingual Graph Label Propagation



Bilingual Graph-Based Label Propagation



End result : larger dictionary



Das

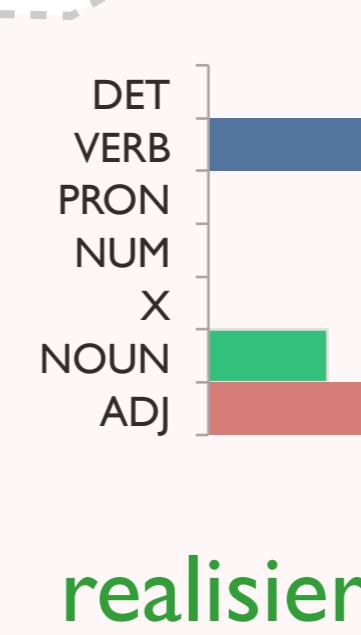
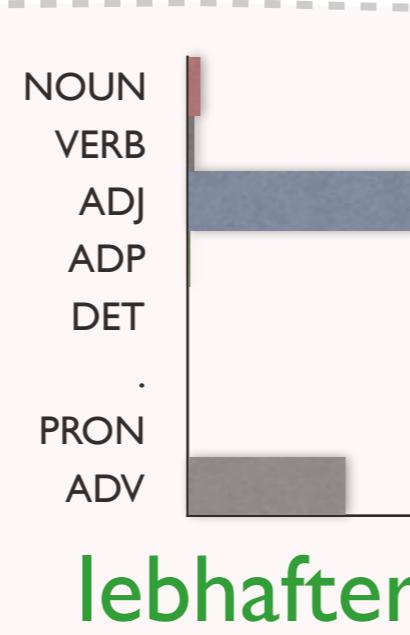
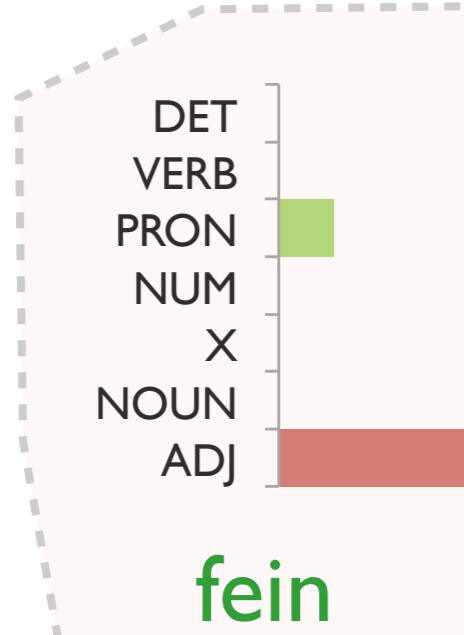
Essen

ist

gut

bei

Google



fein

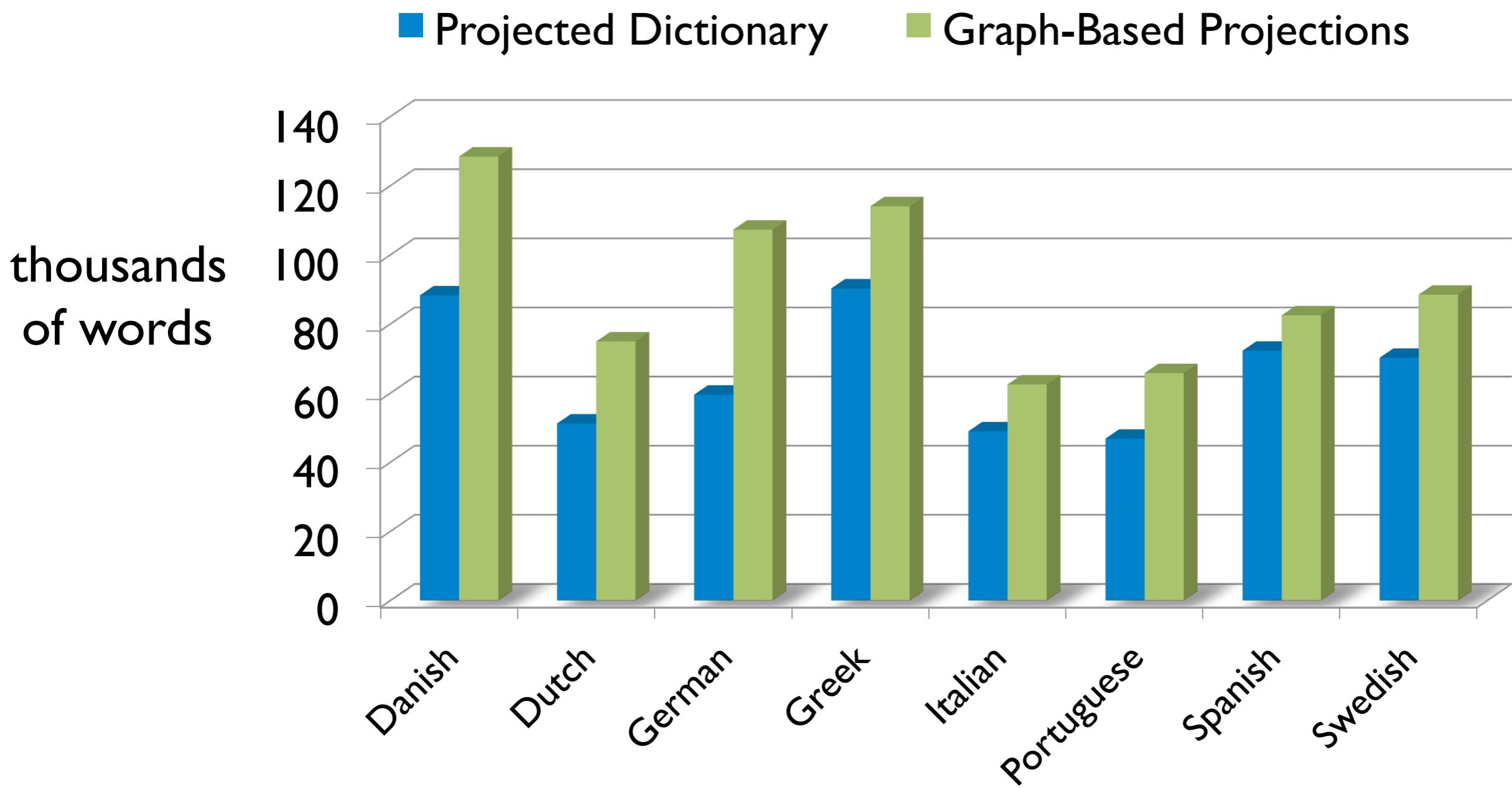
lebhafter

realisieren

Bilingual Graph-Based Label Propagation



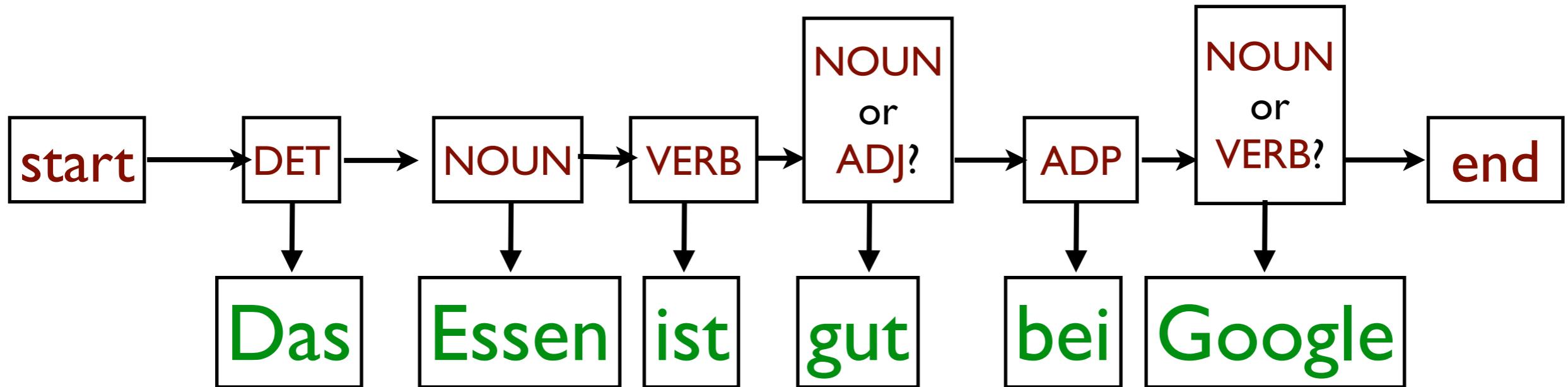
Lexicon Expansion



Unsupervised POS Tagging with **Graph-Based Projections**



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



Model 3 : Graph-Based Lexicon Projection

Unsupervised POS Tagging with **Graph-Based Projections**



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

Unsupervised POS Tagging with **Graph-Based Projections**



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

Model 3 : Graph-Based Lexicon Projection

									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

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences
and Dictionaries

Generalizing above with *Posterior Regularization*

Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences
and Dictionaries

Generalizing above with *Posterior Regularization*

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

- 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

The screenshot shows the Wiktionary entry for the word "Google". At the top, there are tabs for "Eintrag" and "Diskussion". Below the tabs, the word "Google" is displayed in a large, bold, black font. Underneath the word, there is a link to "Siehe auch: google". The main content area is divided into several sections: "Google (Deutsch)" with a "Bearbeiten" link, "Substantiv, Eigenname" with a "Bearbeiten" link, "Worttrennung:" followed by "Goo·gle, kein Plural", "Aussprache:" followed by IPA: ['gu:g!'] and a "Hörbeispiele" link, "Bedeutungen:" followed by "[1] Informationstechnologie: eine Suchmaschine im World Wide Web", and "Herkunft:" followed by a detailed explanation about its etymology. At the bottom, a note states: "Das Wort ist markenschutzrechtlich umkämpft. Siehe auch bei googeln."

Better Dictionaries



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

- Wiktionary

The screenshot shows a Wiktionary entry for the word "NOUN". A speech bubble on the left contains the word "NOUN". A line connects the word "NOUN" to the word "Substantiv, Eigenname" in the Wiktionary entry, which is highlighted with a red oval. The Wiktionary page includes sections for "Google", "Google (Deutsch)", "Substantiv, Eigenname", "Worttrennung", "Aussprache", "Bedeutungen", and "Herkunft".

Eintrag Diskussion

Google

Siehe auch: [google](#)

Google (Deutsch) [Bearbeiten]

Substantiv, Eigenname [Bearbeiten]

Worttrennung:
Goo·gle, kein Plural

Aussprache:
IPA: ['gu:gə]
Hörbeispiele: [Google](#) (Info)

Bedeutungen:
[1] [Informationstechnologie](#): eine Suchmaschine im World Wide Web

Herkunft:
abgeleitet von der amerikanischen Aussprache der Zahl „googol“ → en
jeglichen ethymologischen Hintergrund erfunden wurde; die Betreiber d
widerspiegeln soll.

Das Wort ist markenschutzrechtlich umkämpft. Siehe auch bei [googeln](#).

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

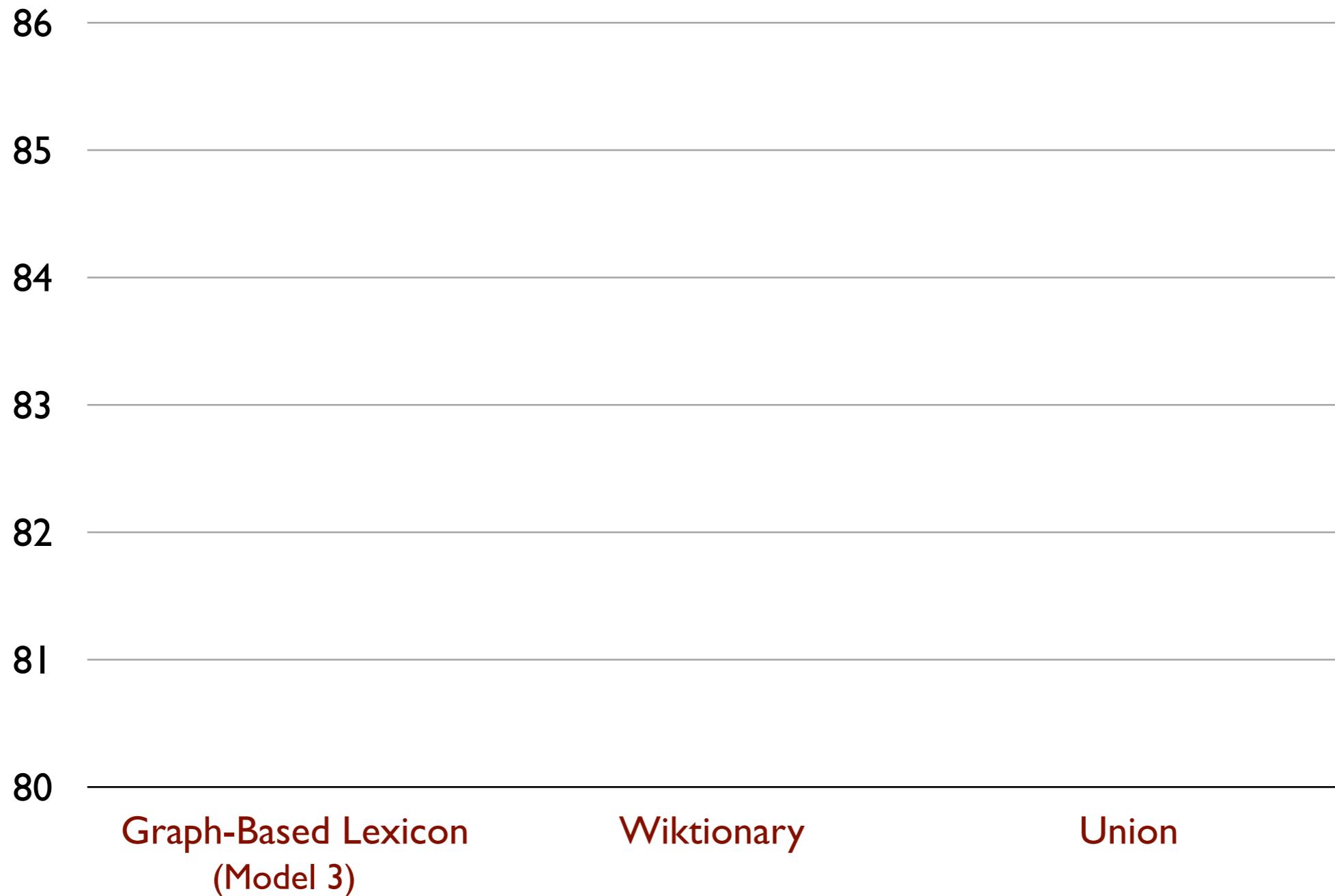


Measured across 8 Indo-European Languages

HMM with Wiktionary Constraints



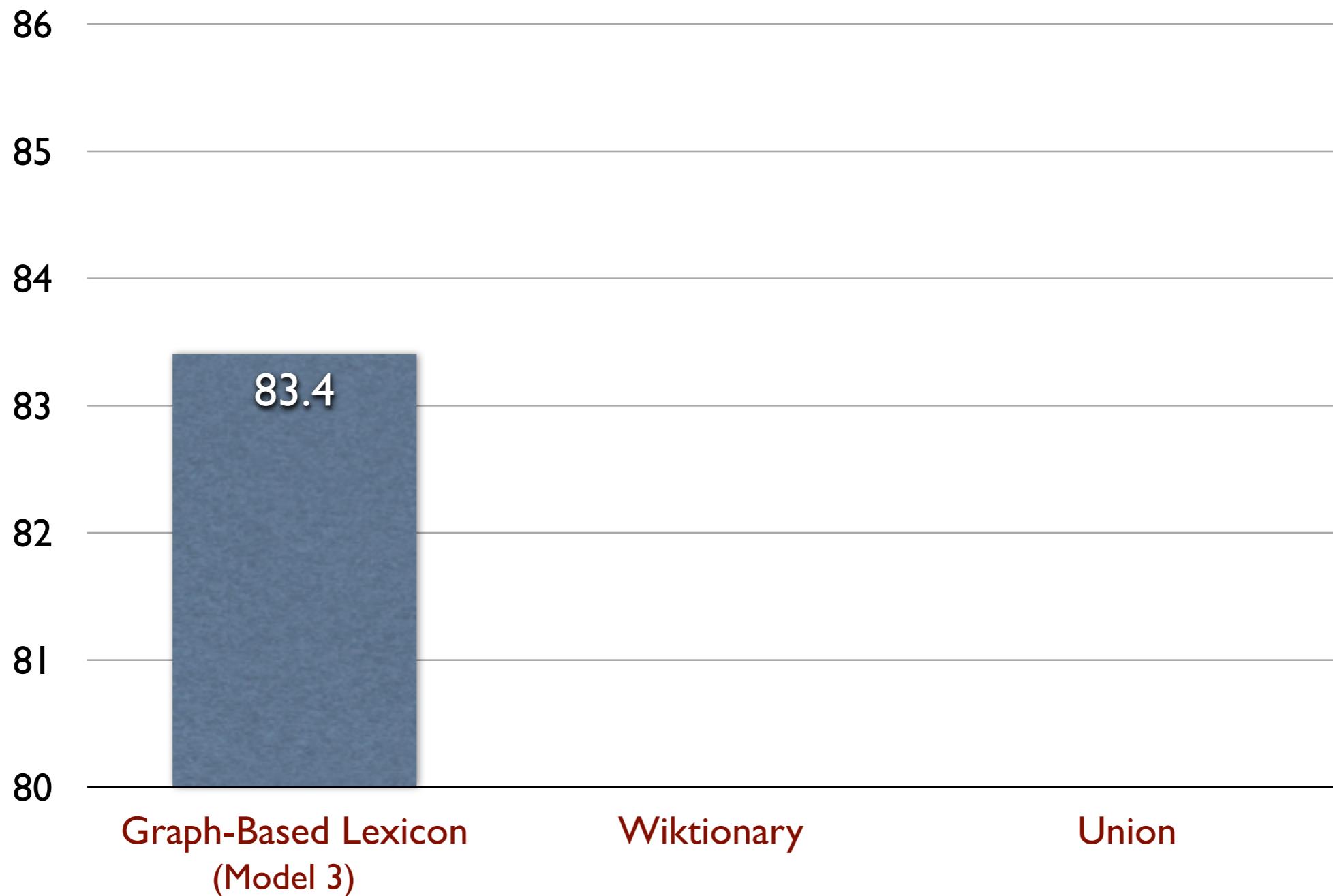
Measured across 8 Indo-European Languages



HMM with Wiktionary Constraints



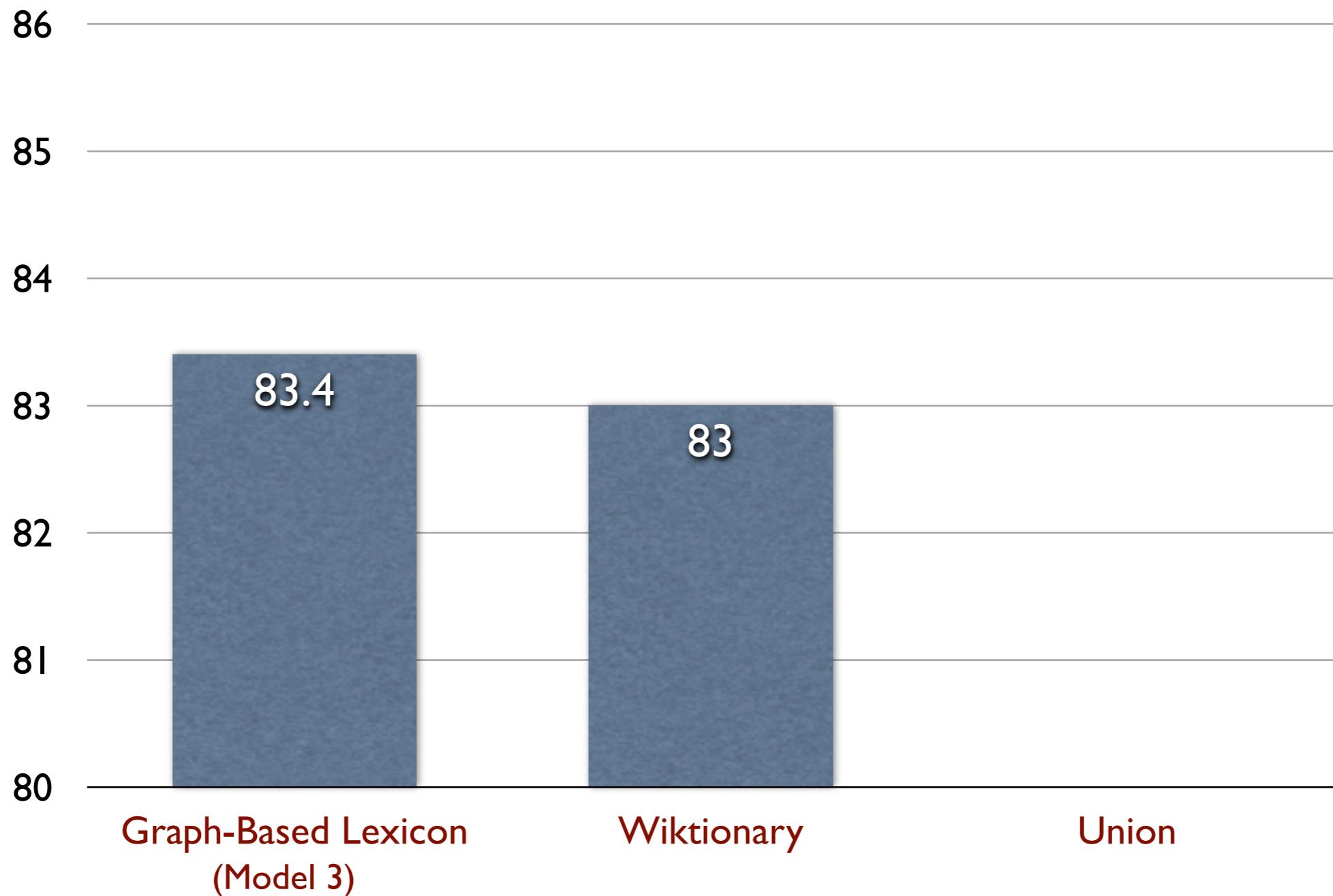
Measured across 8 Indo-European Languages



HMM with Wiktionary Constraints



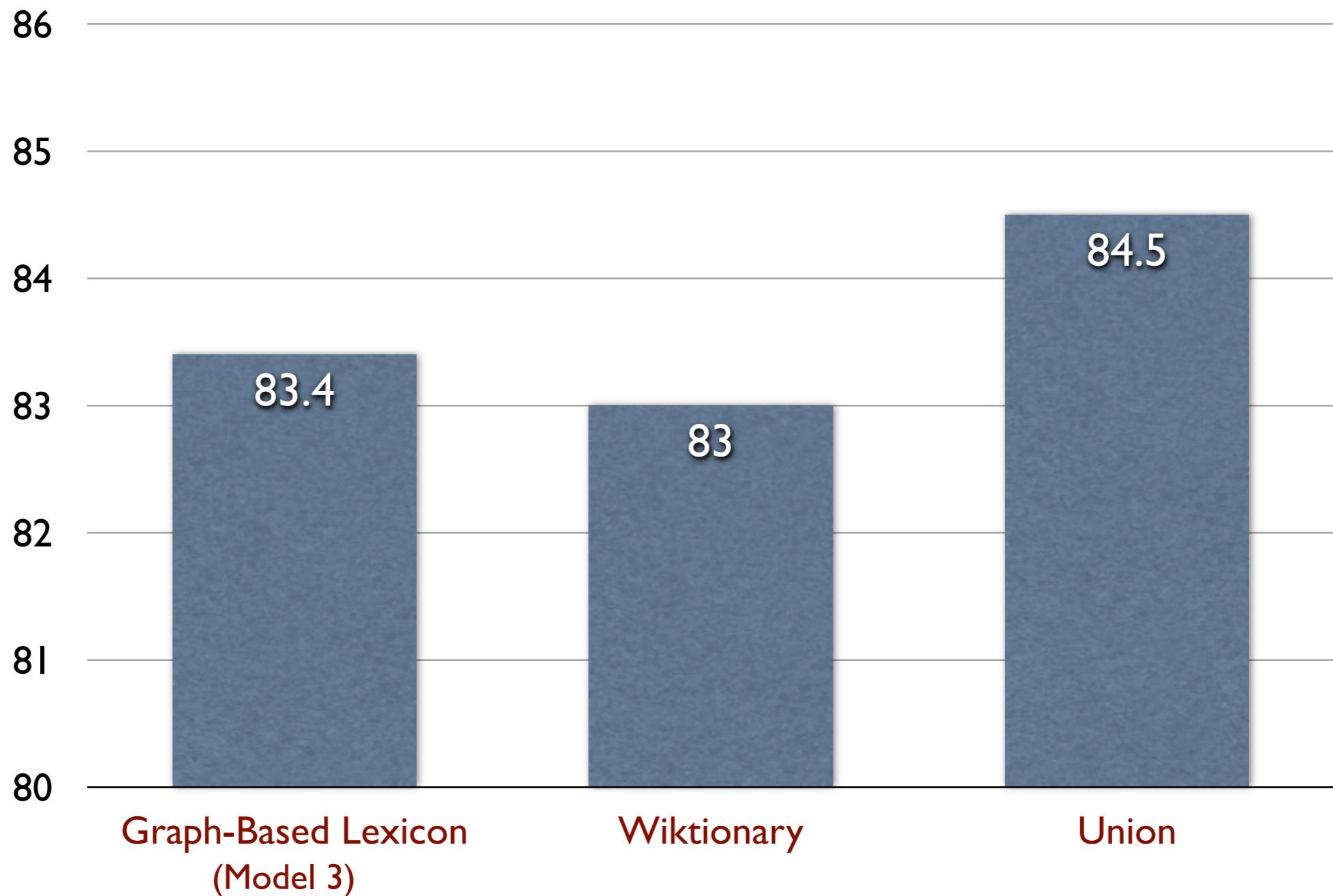
Measured across 8 Indo-European Languages



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

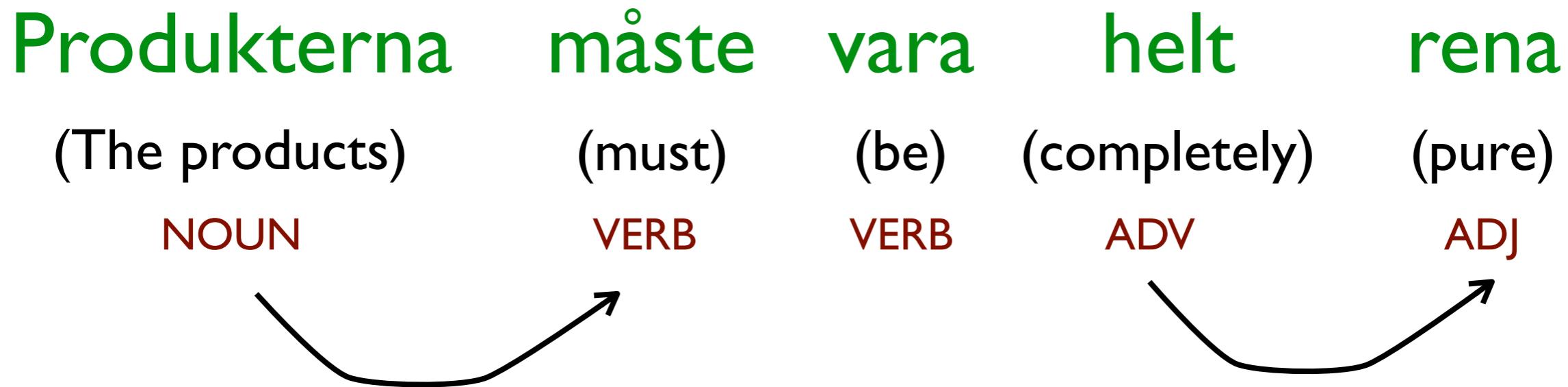


Transition information in provided supervision

Discriminative model (Conditional Random Field):

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

Complete Token Supervision



Transition information in provided supervision

Discriminative model (Conditional Random Field):

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

All possible tag sequences for 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

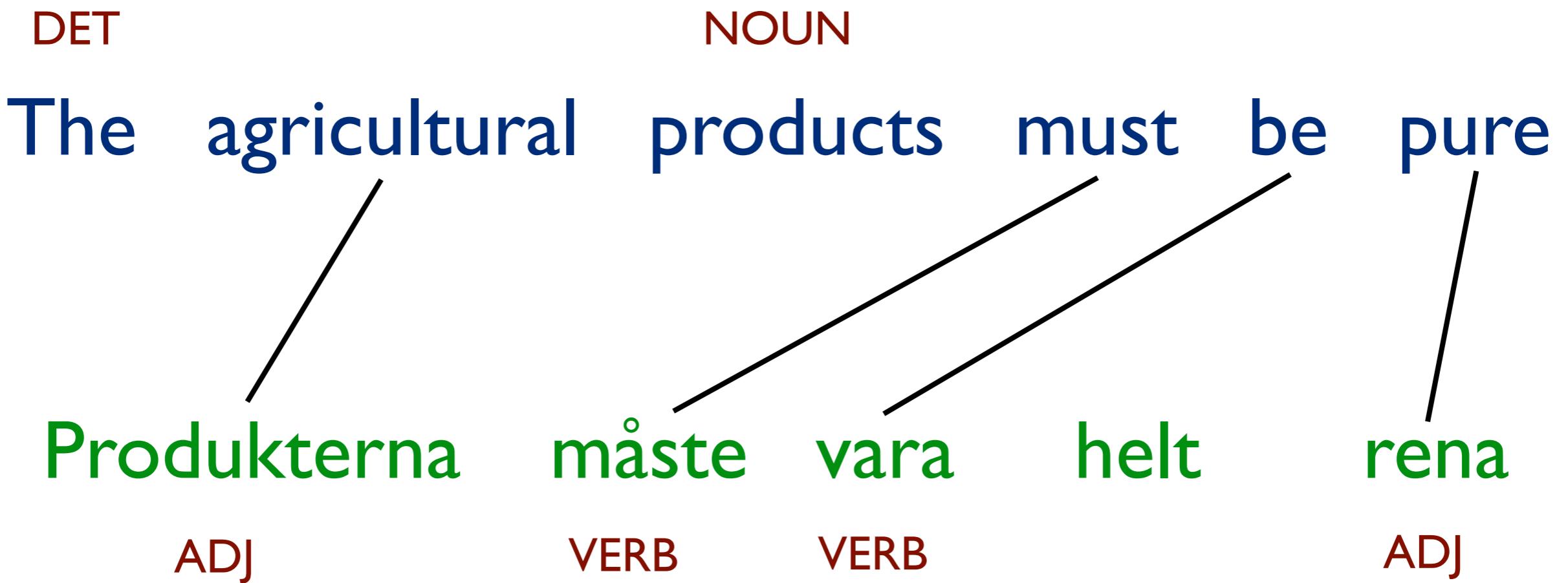
Parallel Sentence Pair



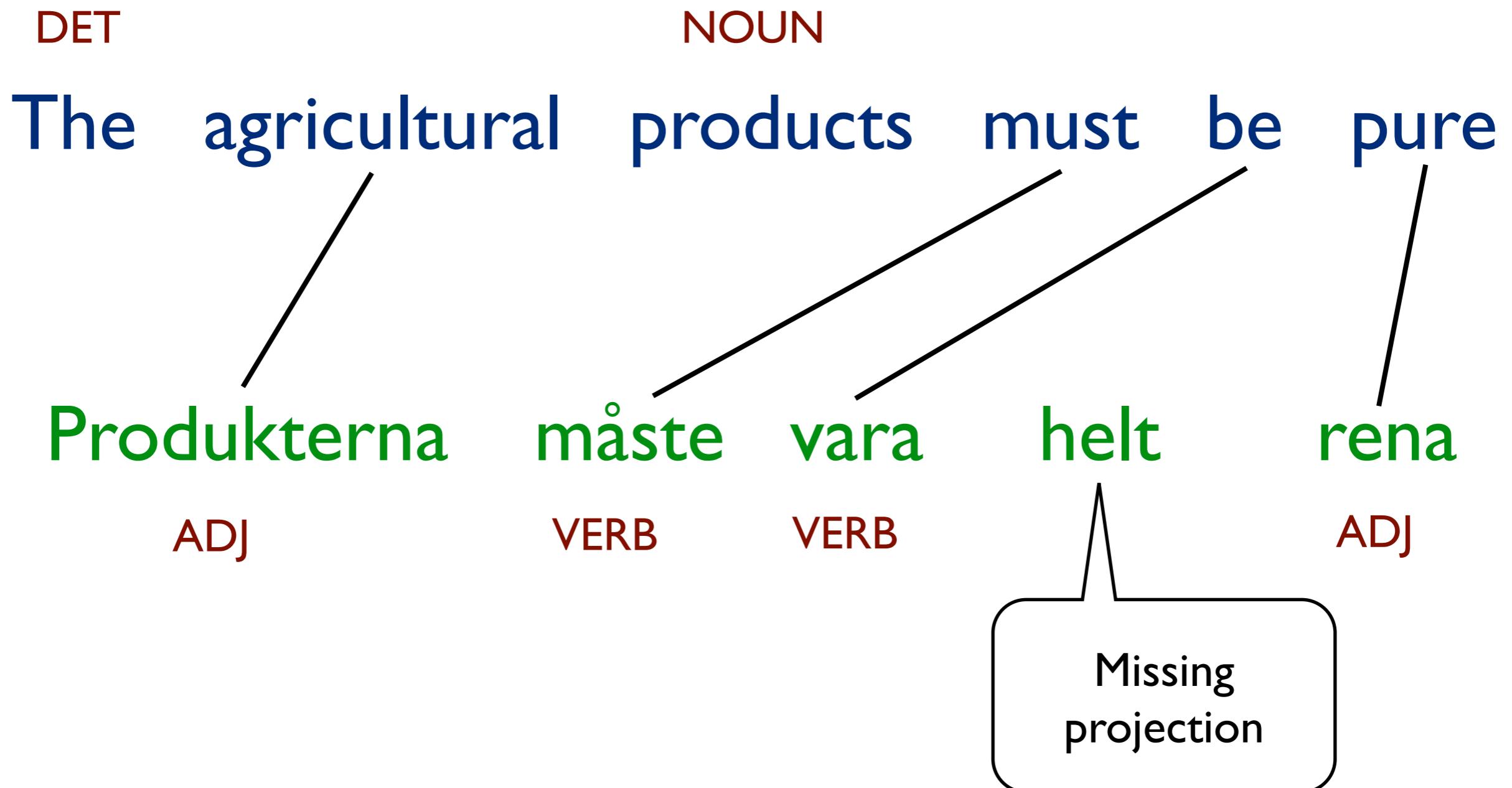
DET	ADJ	NOUN	VERB	VERB	ADJ
The	agricultural	products	must	be	pure
Produkterna	måste	vara	helt		rena

The diagram illustrates a parallel sentence pair between English and Swedish. The English sentence is "The agricultural products must be pure". The Swedish sentence is "Produkterna måste vara helt rena". Part-of-speech tags are shown above the English words: DET, ADJ, NOUN, VERB, VERB, and ADJ. Lines connect the English words to their corresponding Swedish words: 'The' connects to 'Produkterna', 'agricultural' connects to 'måste', 'products' connects to 'vara', 'must' connects to 'helt', 'be' connects to an empty space, and 'pure' connects to 'rena'. The Swedish words 'måste' and 'vara' are both labeled as VERB, while 'helt' and 'rena' are labeled as ADJ.

Parallel Sentence Pair



Parallel Sentence Pair



Parallel Sentence Pair

The Google logo is displayed in its signature multi-colored font. The letters are arranged in a bold, rounded style. The colors used are blue for 'G', red for 'o', orange for 'o', green for 'g', blue for 'l', and red for 'e'. A small 'TM' symbol is located at the top right corner of the 'e'.

The diagram illustrates the morphological structure of the sentence "The agricultural products must be pure". The English words are shown in blue, and the Swedish words are shown in green. Part-of-speech labels (DET, NOUN, ADJ, VERB) are in red, and error annotations are in black.

English Morphology:

- "The" is labeled DET.
- "agricultural" is labeled NOUN.
- "products" is labeled NOUN.
- "must" is labeled VERB.
- "be" is labeled VERB.
- "pure" is labeled ADJ.

Swedish Morphology:

- "Produkterna" is labeled ADJ.
- "måste" is labeled VERB.
- "vara" is labeled VERB.
- "helt" is labeled ADJ.
- "rena" is labeled ADJ.

Annotations:

- A box labeled "Wrong projection (systematic)" points to the English word "agricultural", which is labeled as NOUN but appears to be derived from the adjective "agricultural".
- A box labeled "Missing projection" points to the English word "pure", which is labeled as ADJ but does not appear to have a corresponding adjective form in the Swedish translation.

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

← →

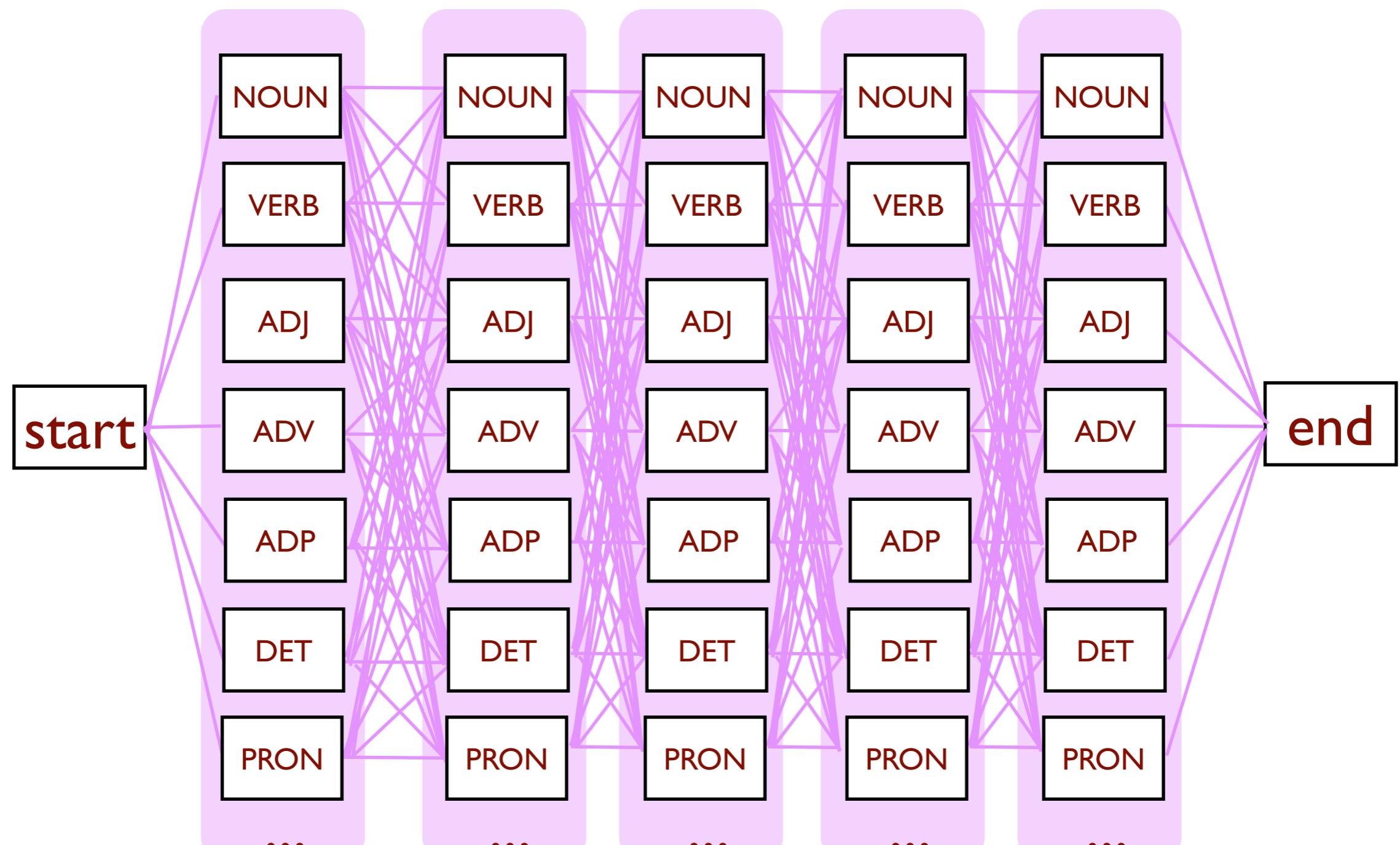
projected tags

Incomplete Token Level Supervision



Produkterna måste vara helt rena

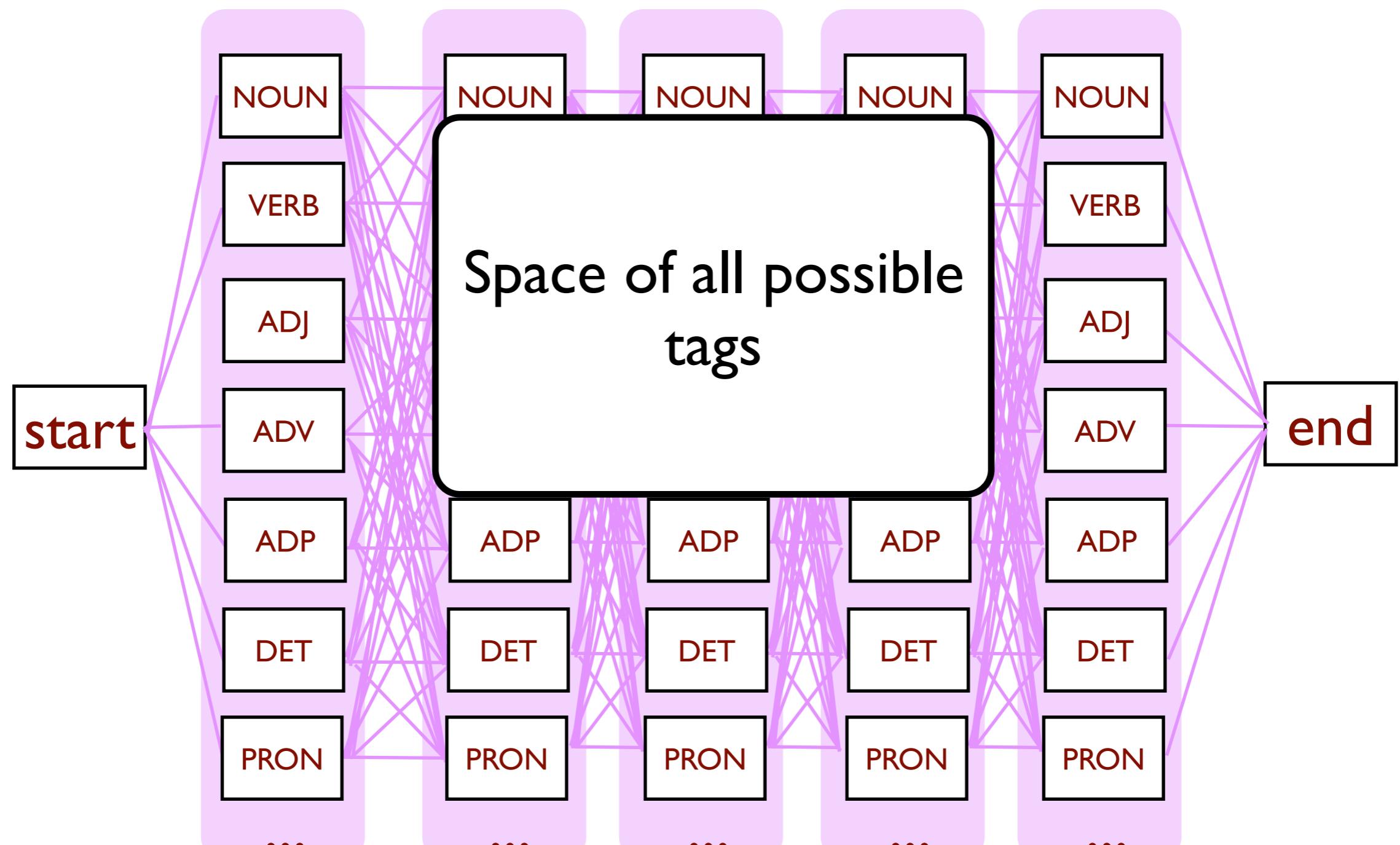
ADJ VERB VERB ADJ



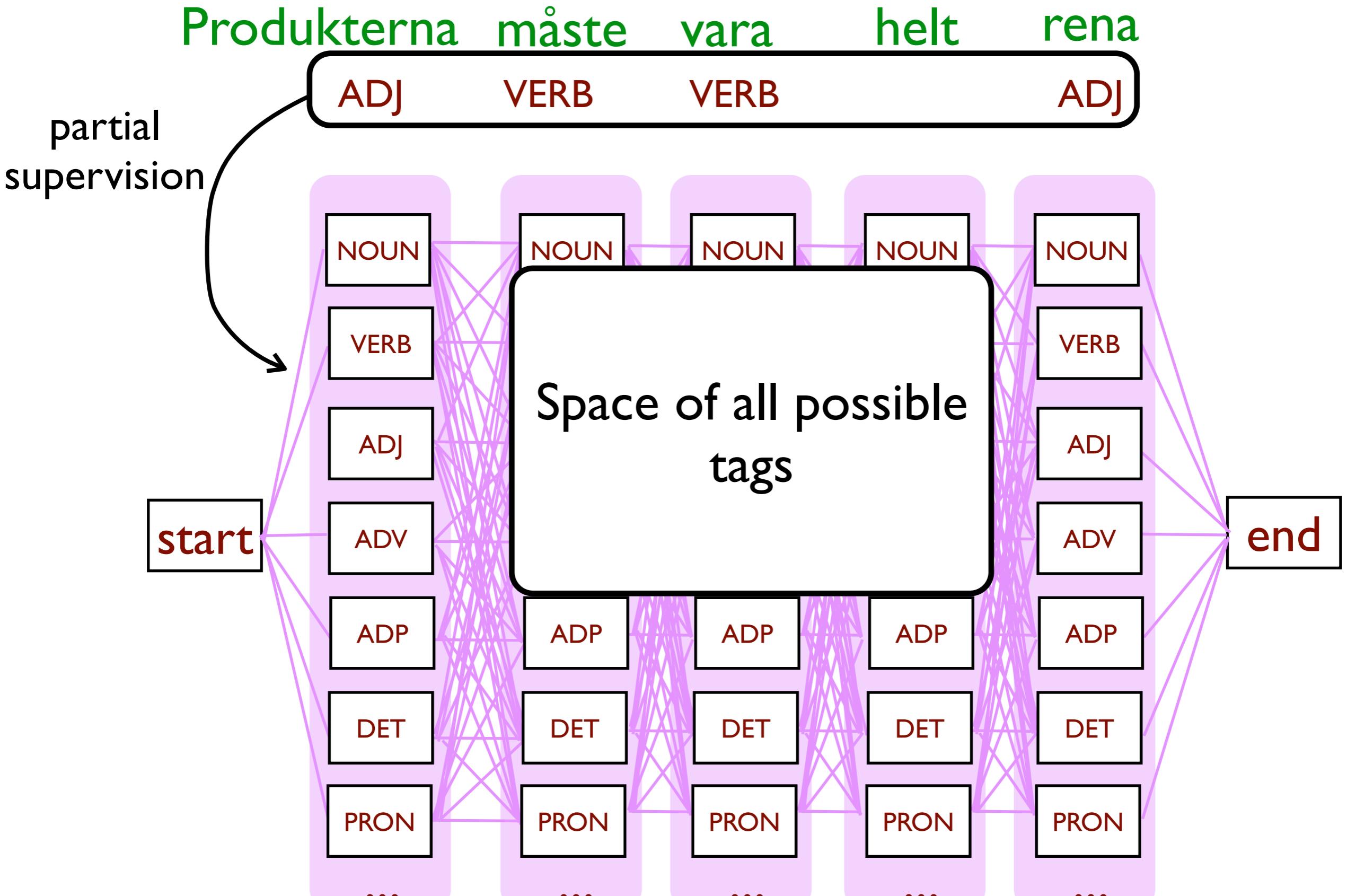
Incomplete Token Level Supervision



Produkterna måste vara helt rena
ADJ VERB VERB ADJ



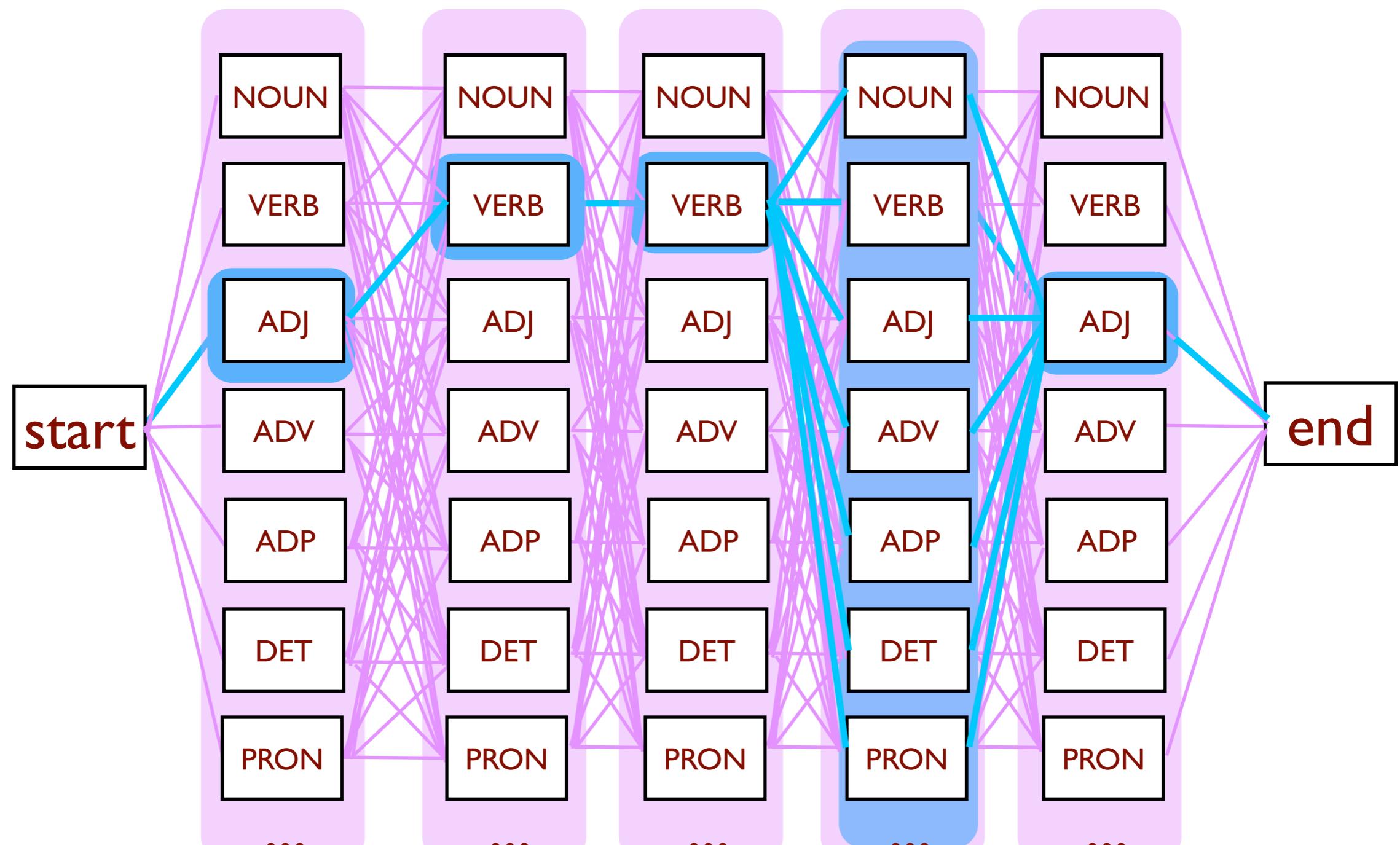
Incomplete Token Level Supervision



Incomplete Token Level Supervision



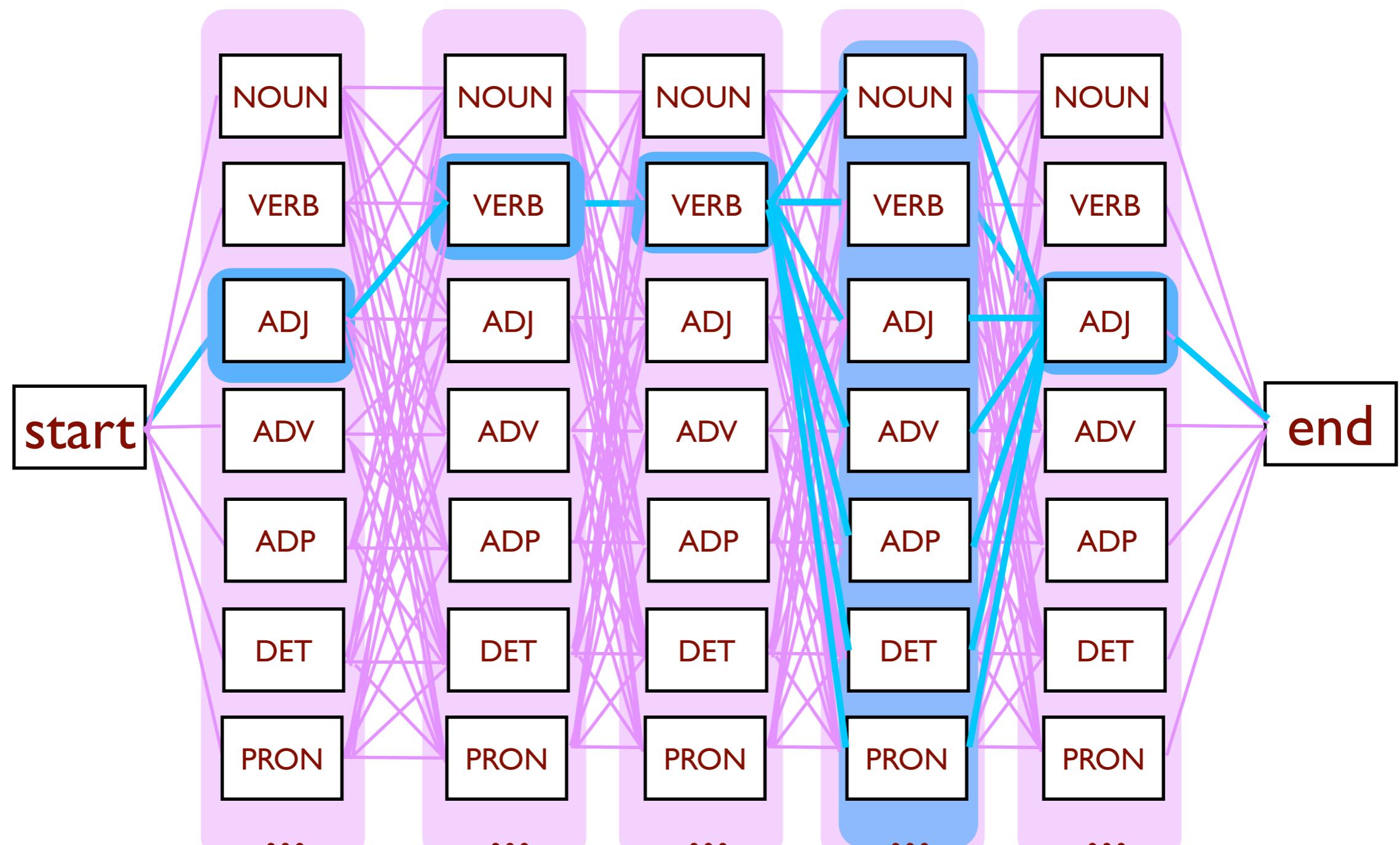
Produkterna måste vara helt rena
ADJ VERB VERB ADJ



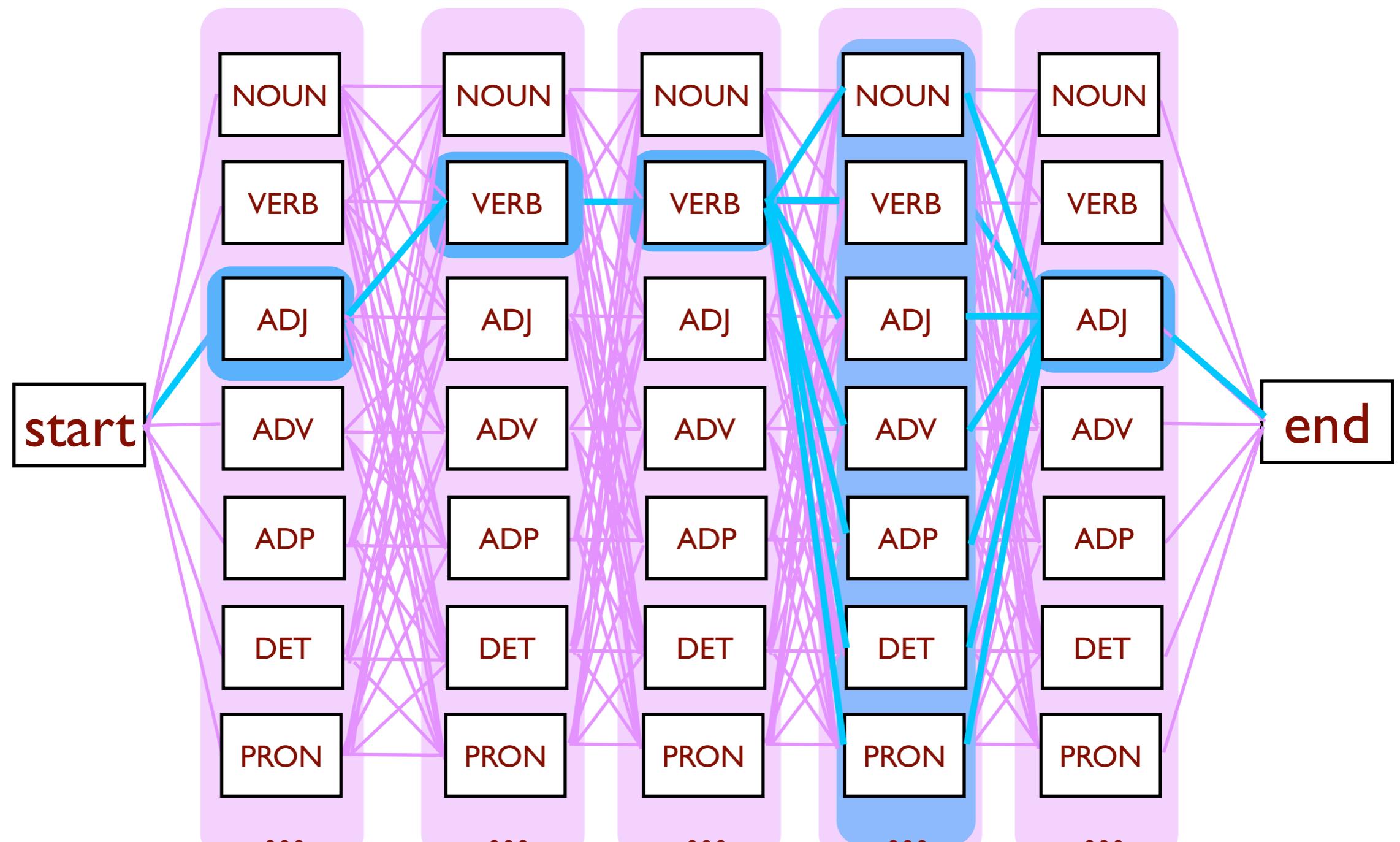
Incomplete Token Level Supervision



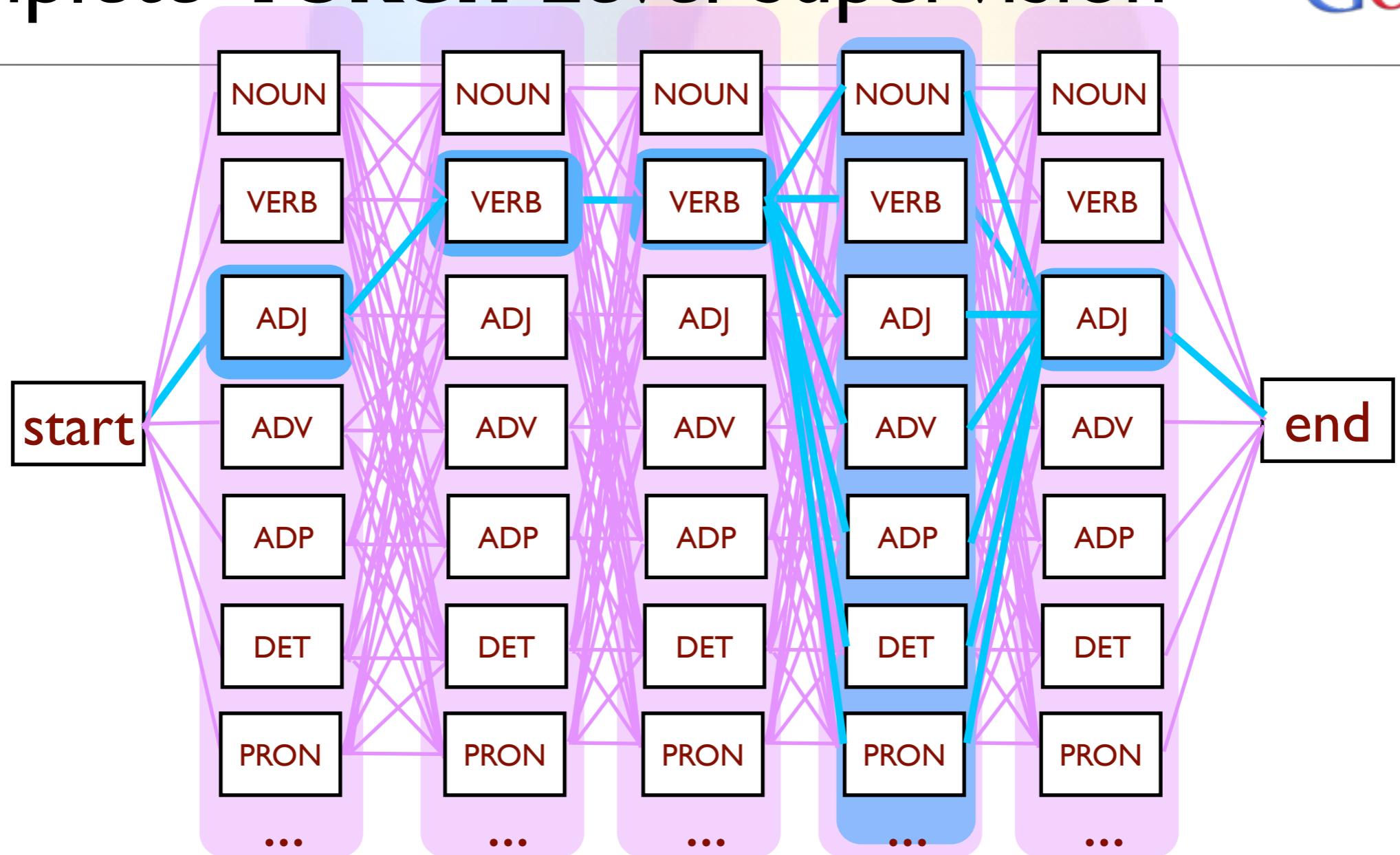
Produkterna måste vara helt rena
ADJ VERB VERB ADJ



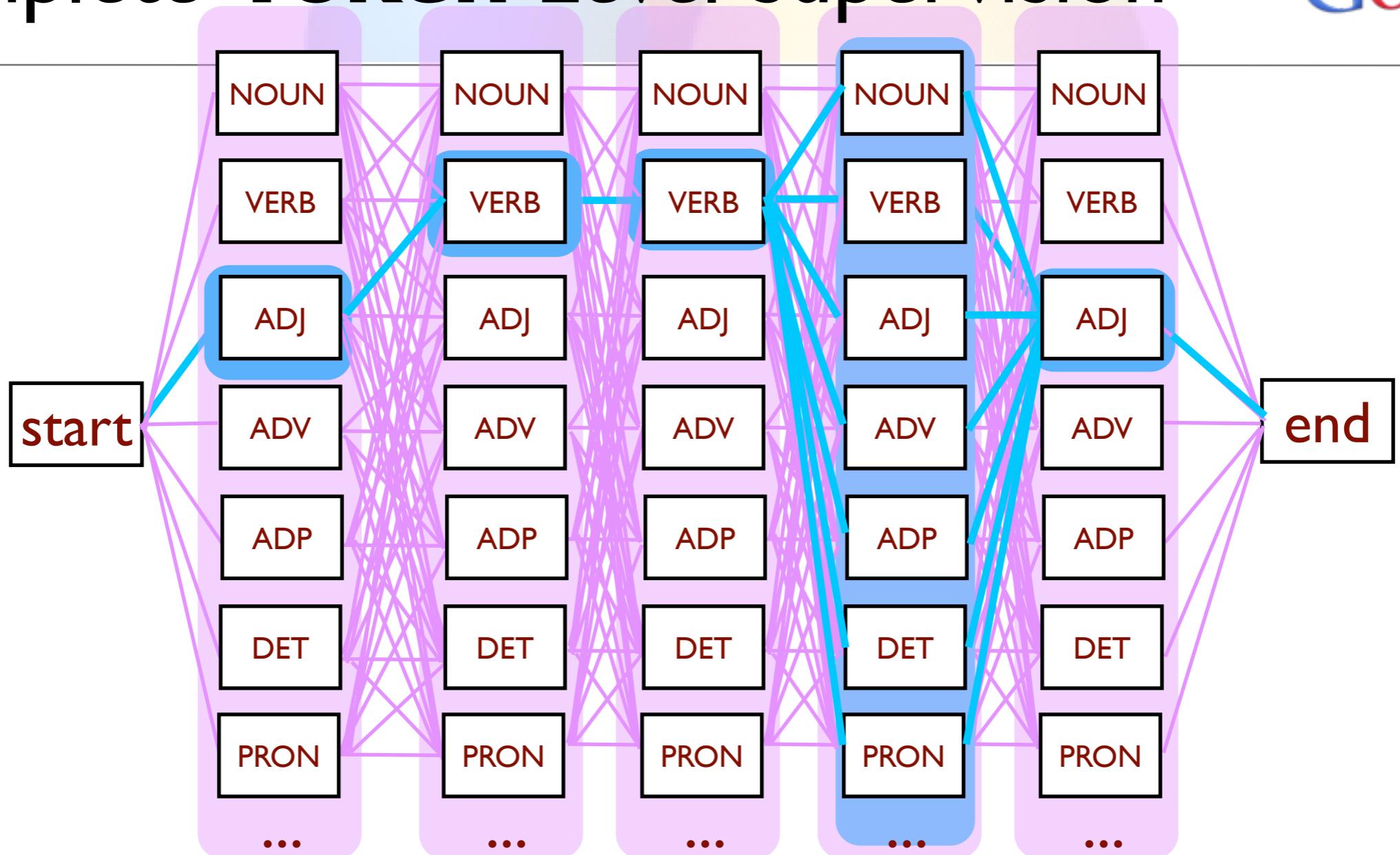
Incomplete Token Level Supervision



Incomplete Token Level Supervision

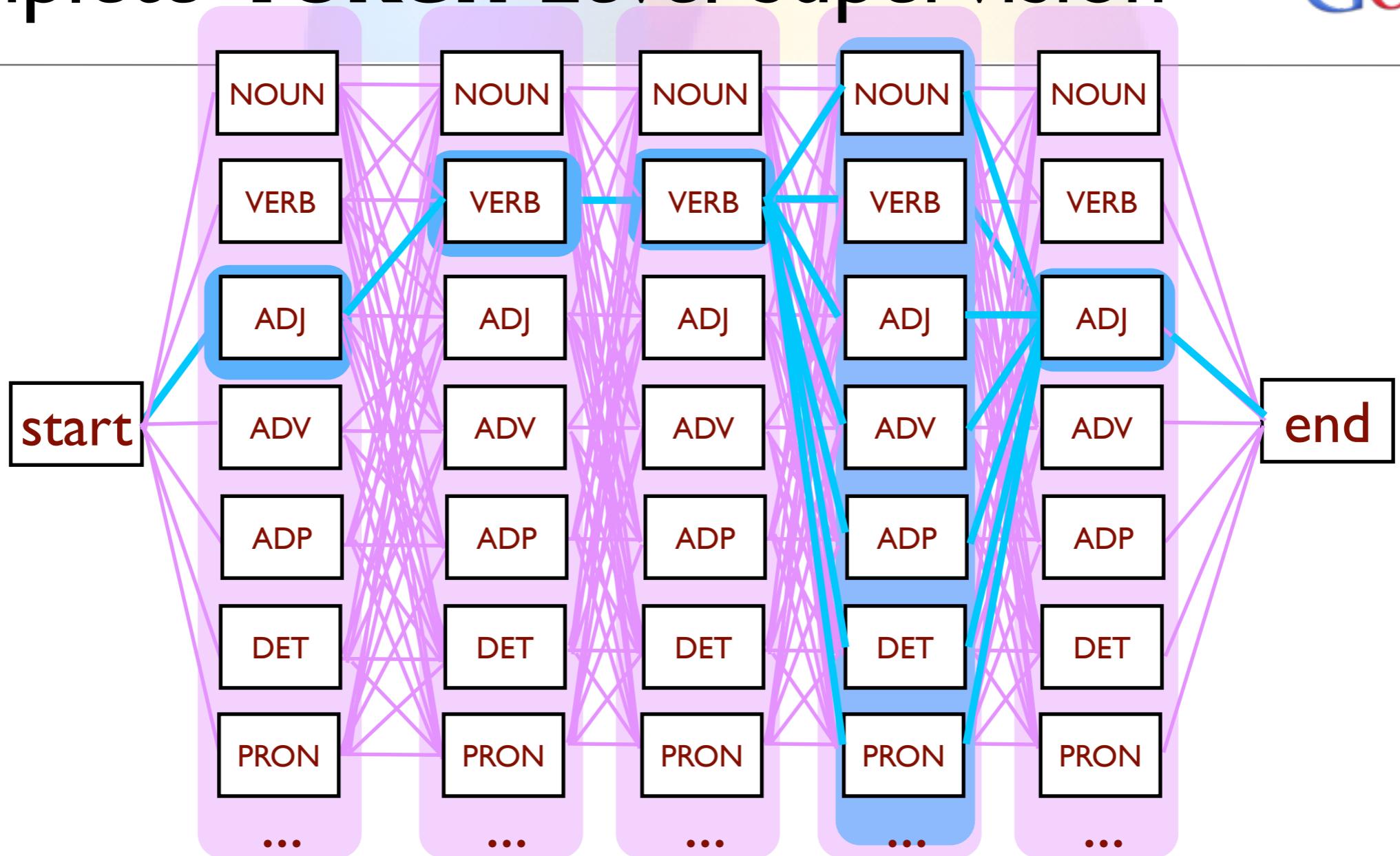


Incomplete Token Level Supervision



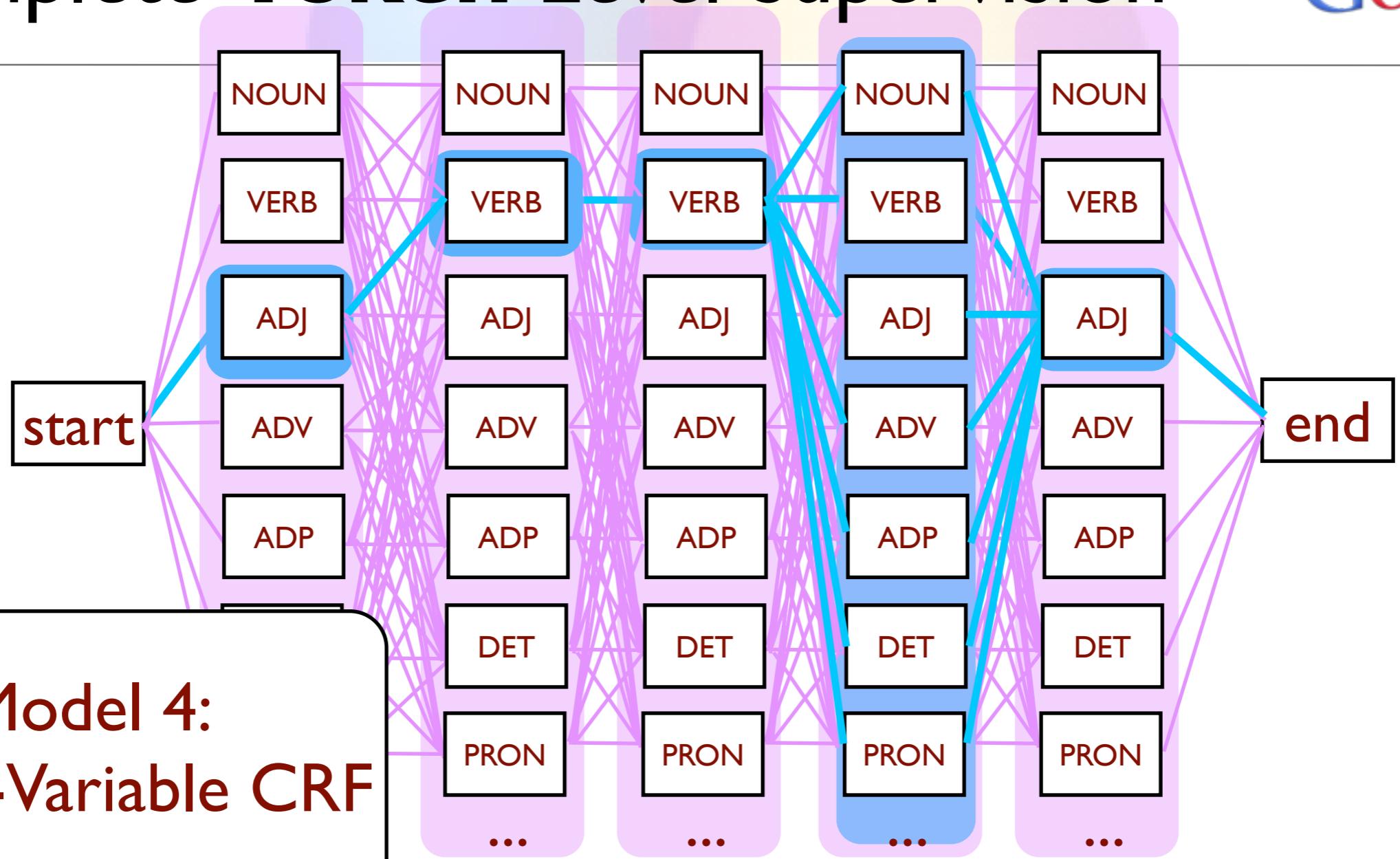
$$p(\tilde{y}|x) = \frac{\sum_{y \in \tilde{y}} \exp \theta \cdot f(x, y)}{\sum_{y' \in \mathcal{Y}(x)} \exp \theta \cdot f(x, y')}$$

Incomplete Token Level Supervision



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

Incomplete Token Level Supervision



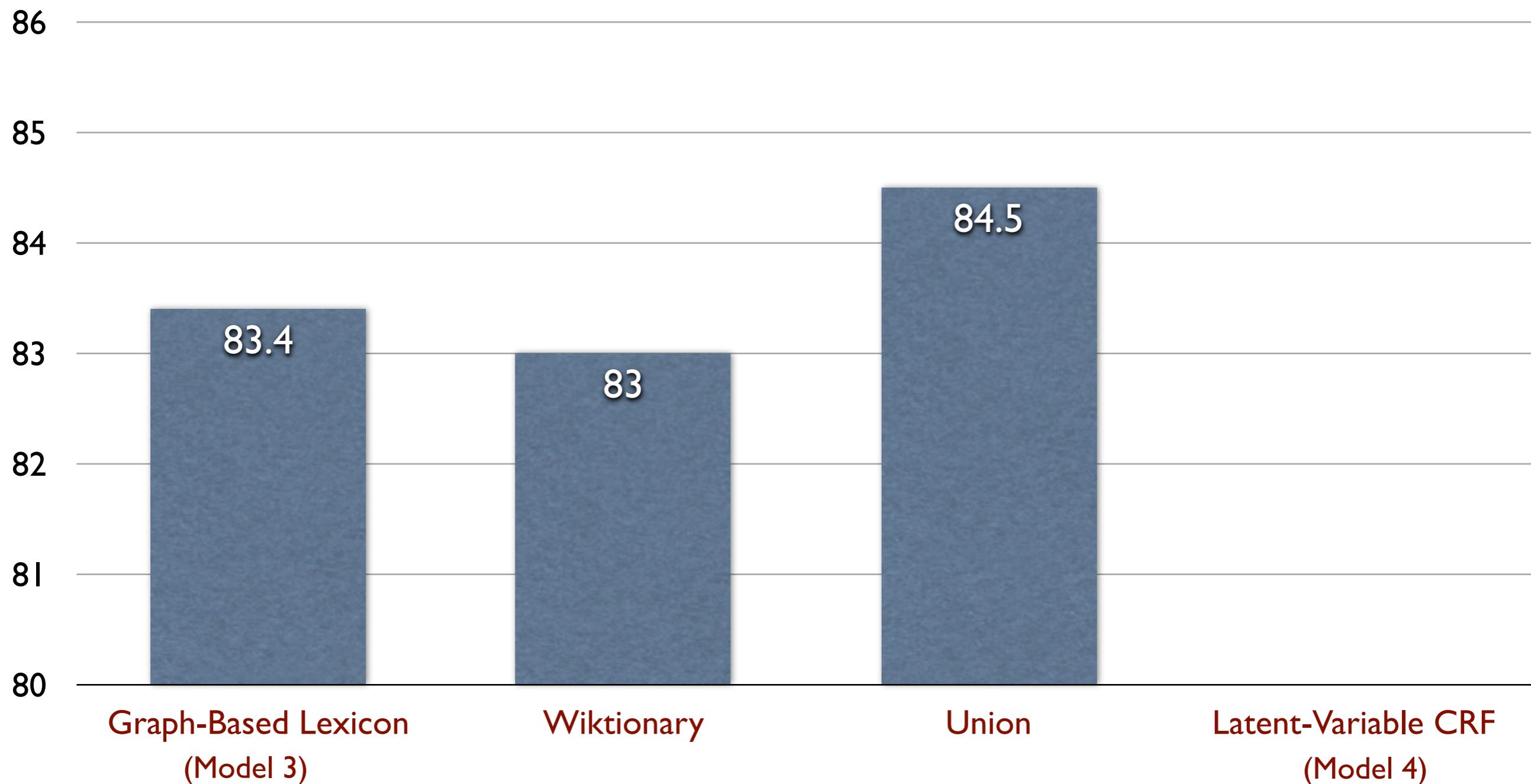
Model 4:
Latent-Variable CRF

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

Model Comparison

Google™

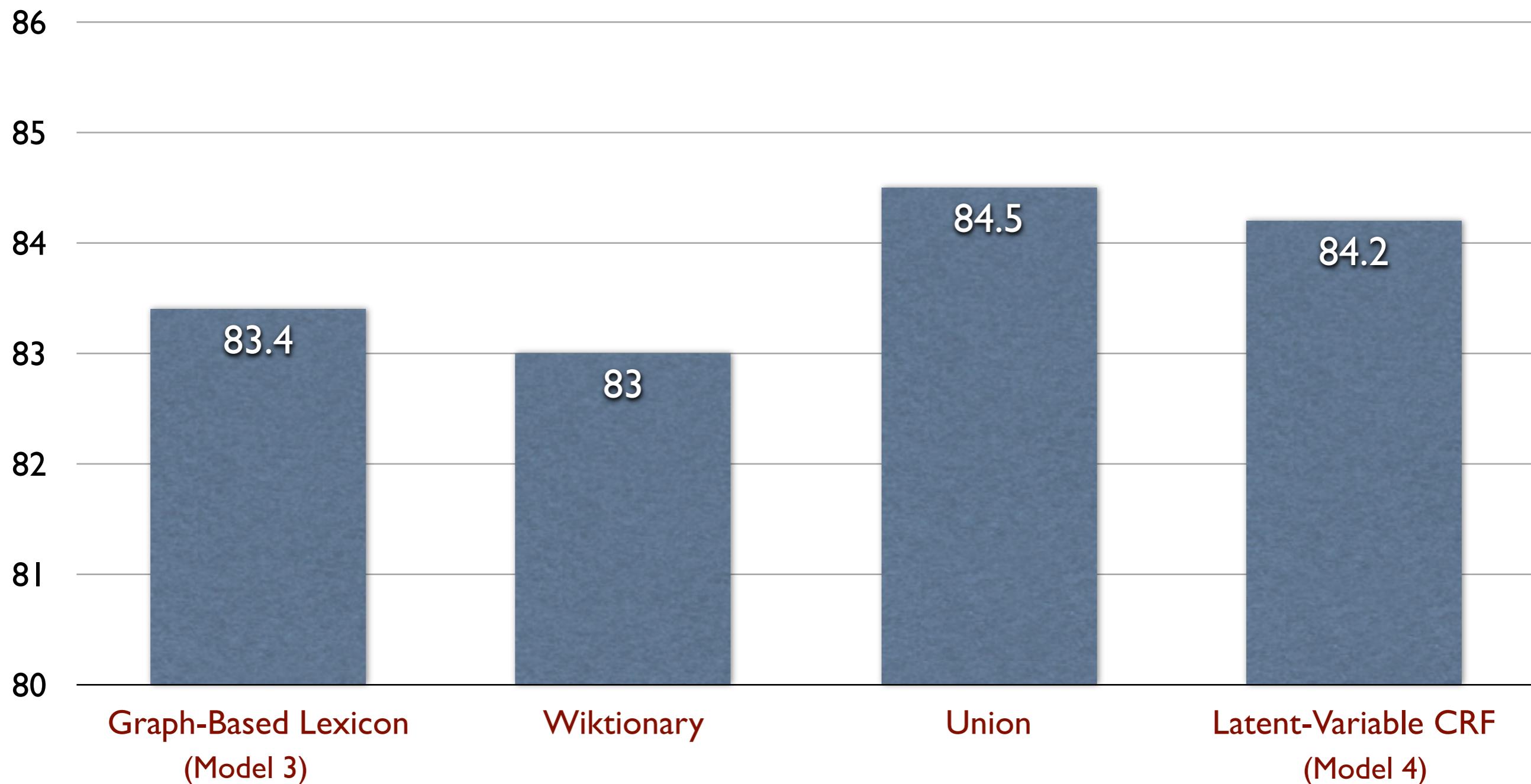
Measured across 8 Indo-European Languages



Model Comparison

Google™

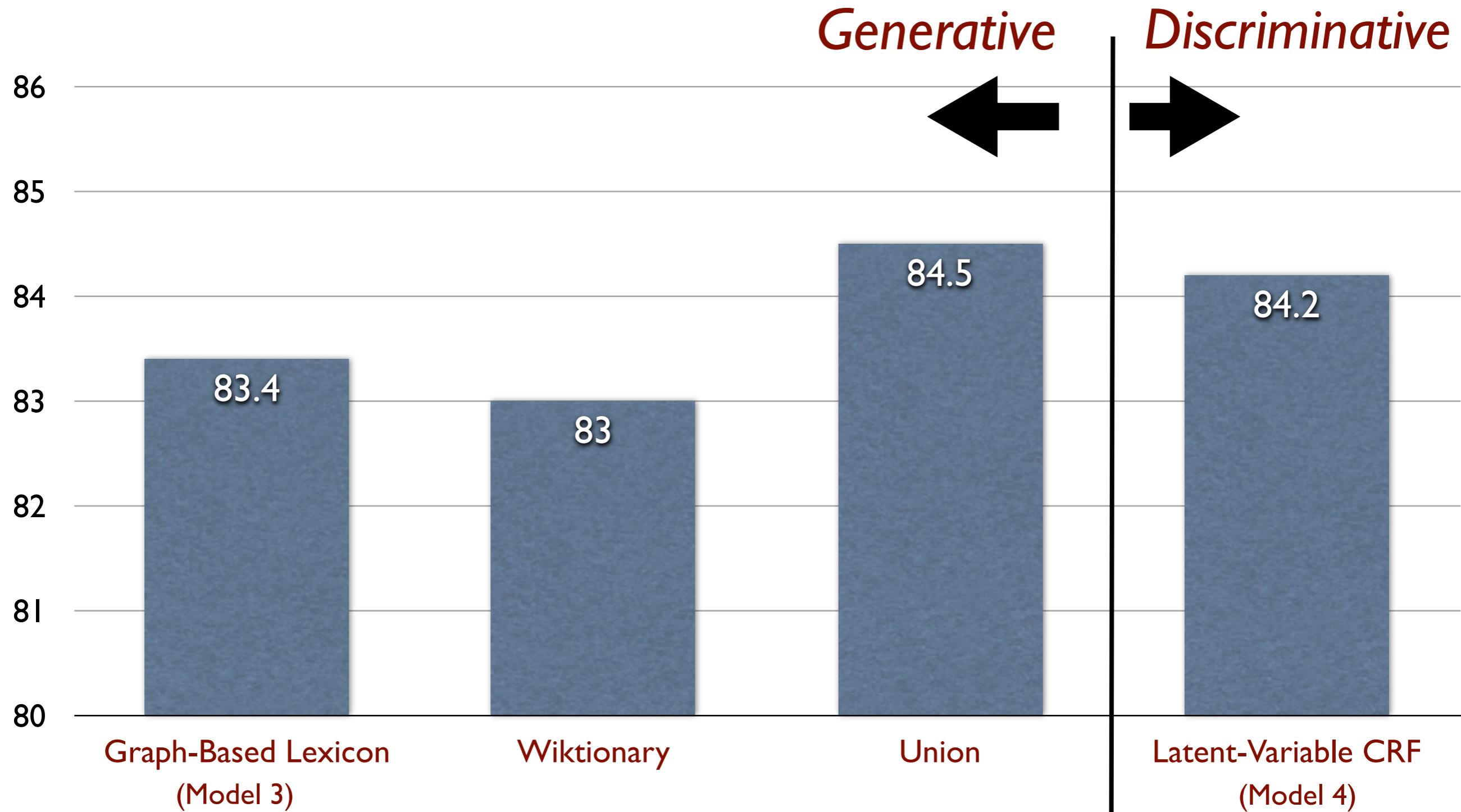
Measured across 8 Indo-European Languages



Model Comparison

Google™

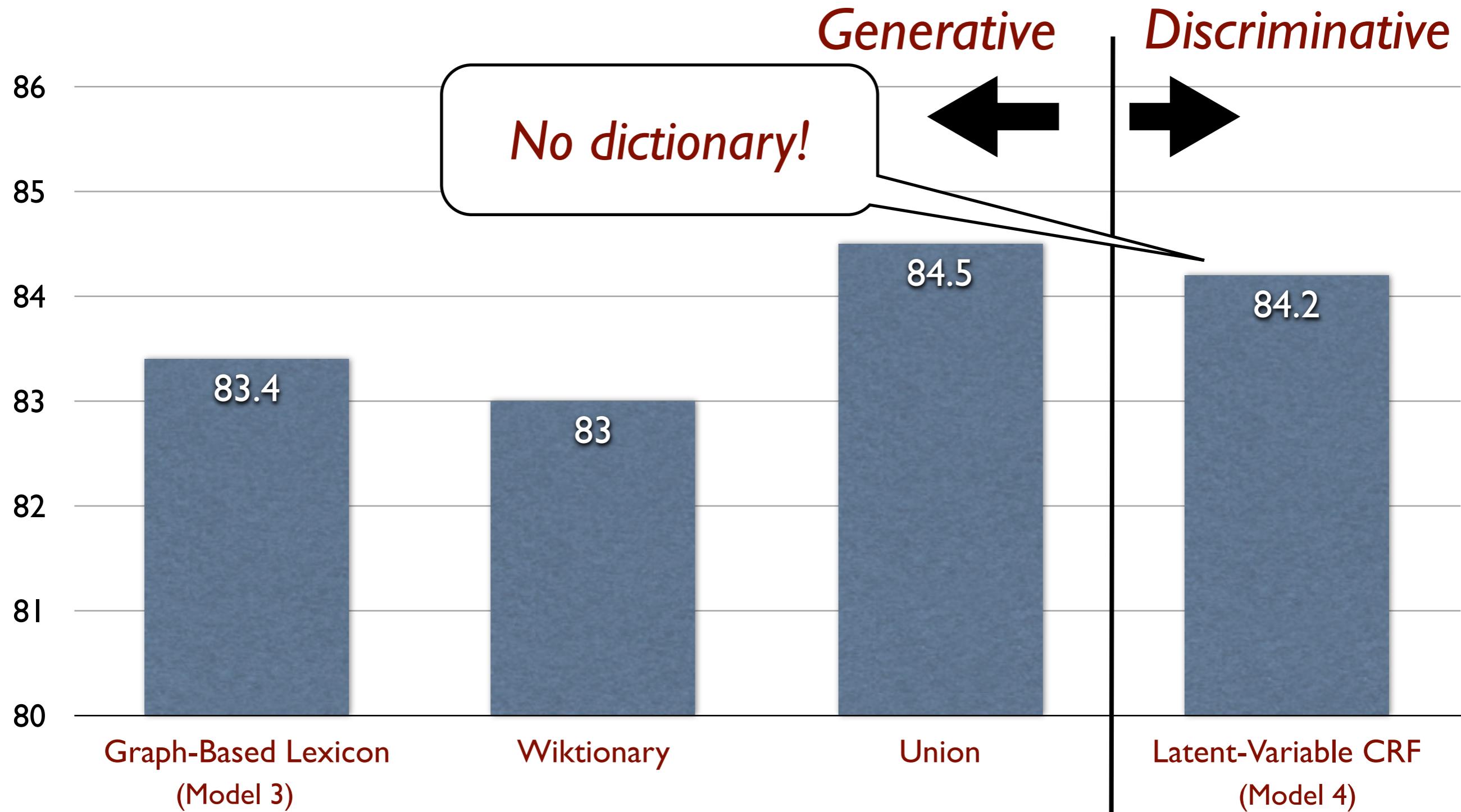
Measured across 8 Indo-European Languages



Model Comparison

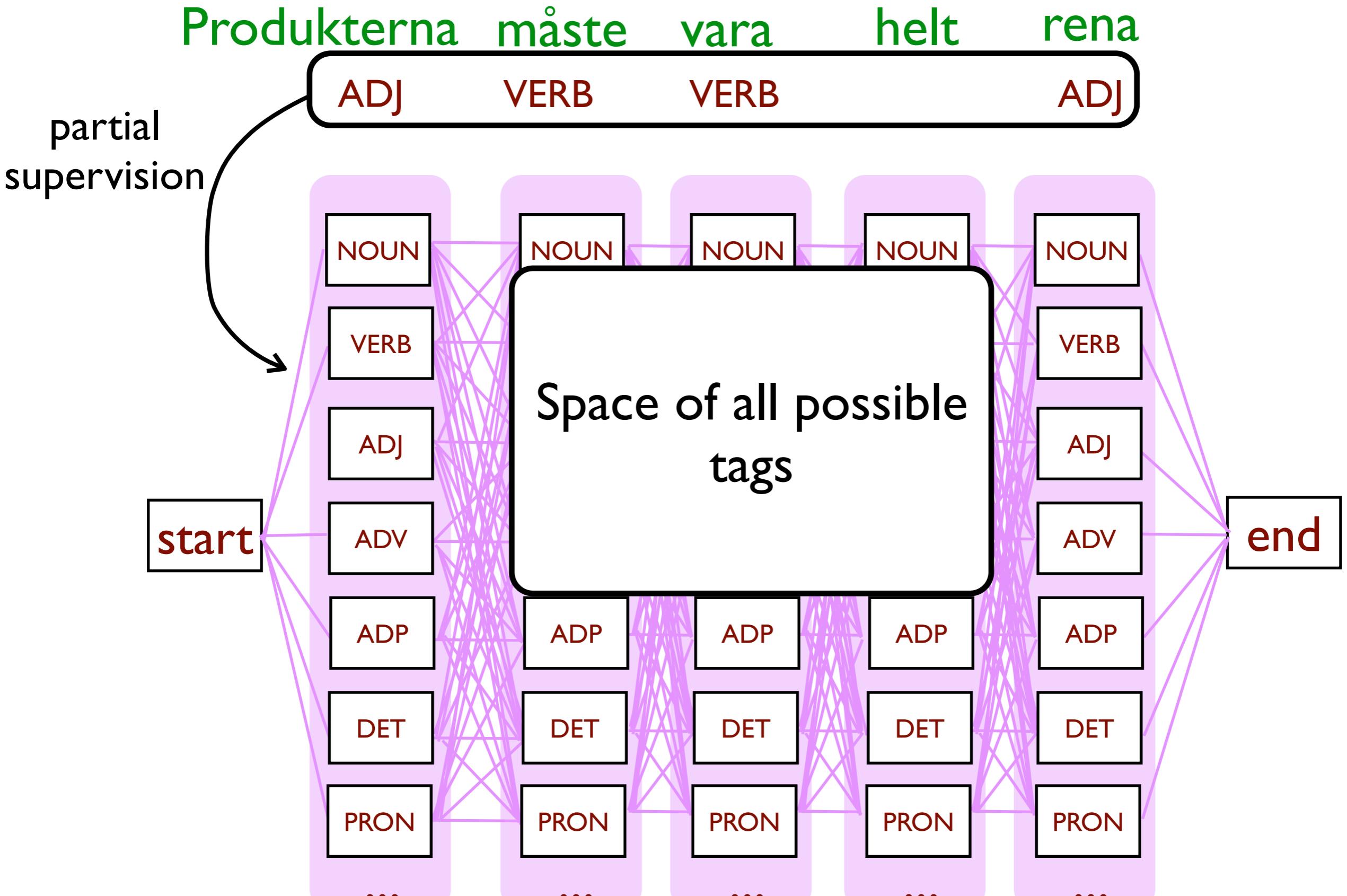


Measured across 8 Indo-European Languages



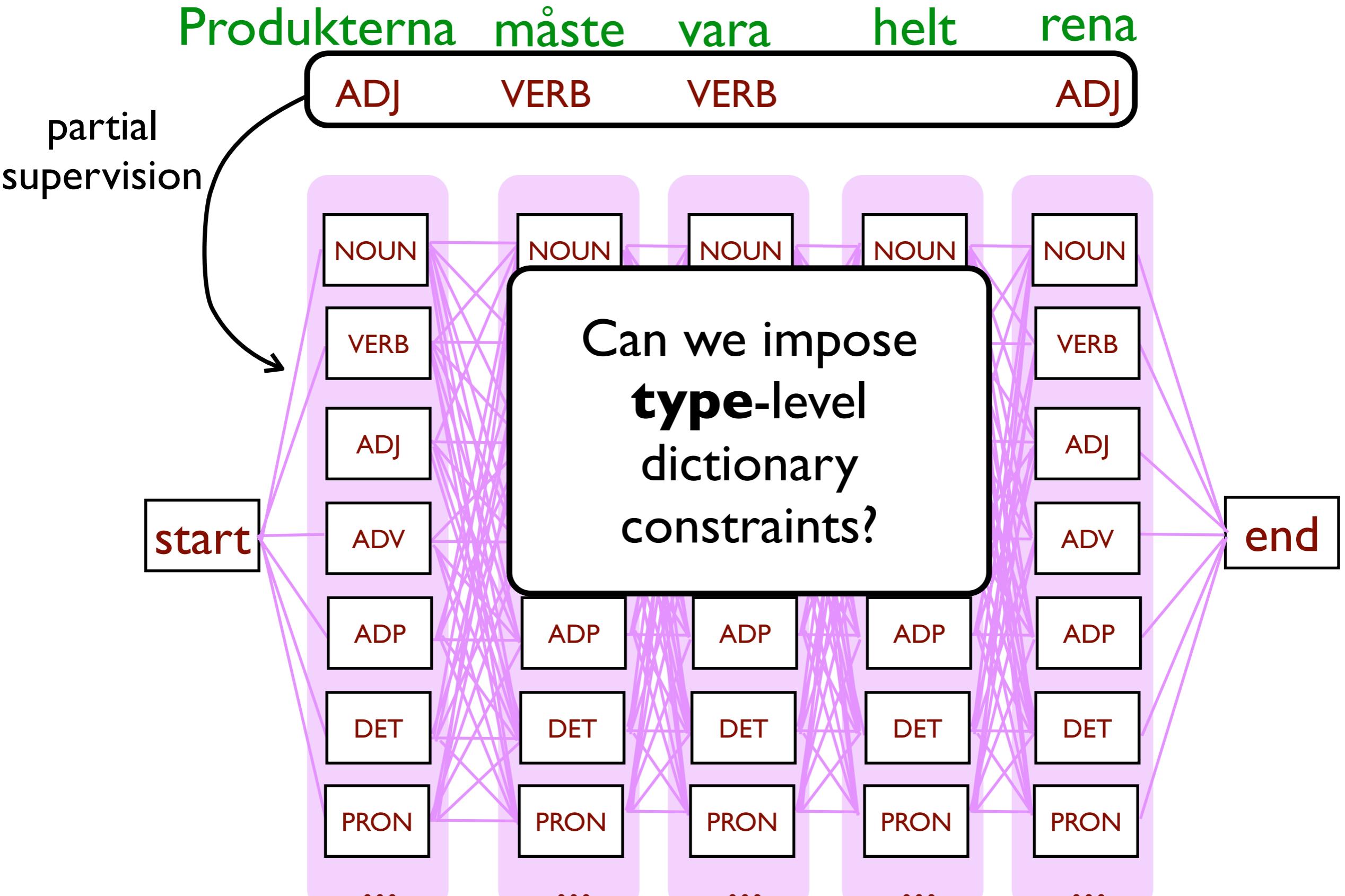
Coupled Token and Type Constraints

Google™



Coupled Token and Type Constraints

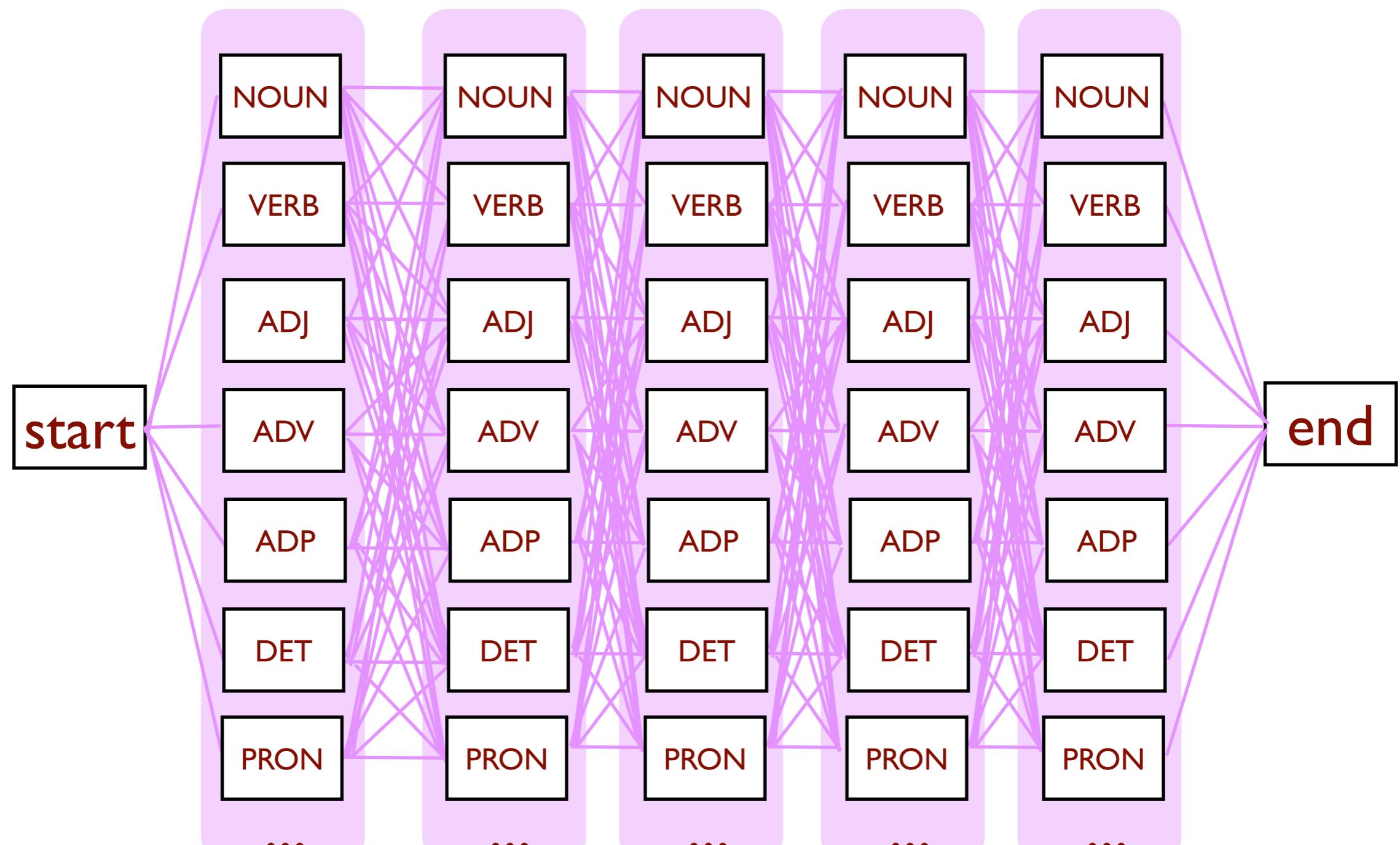
Google™



Coupled Token and Type Constraints

Google™

Produkterna
måste vara helt
ADJ VERB VERB ADJ

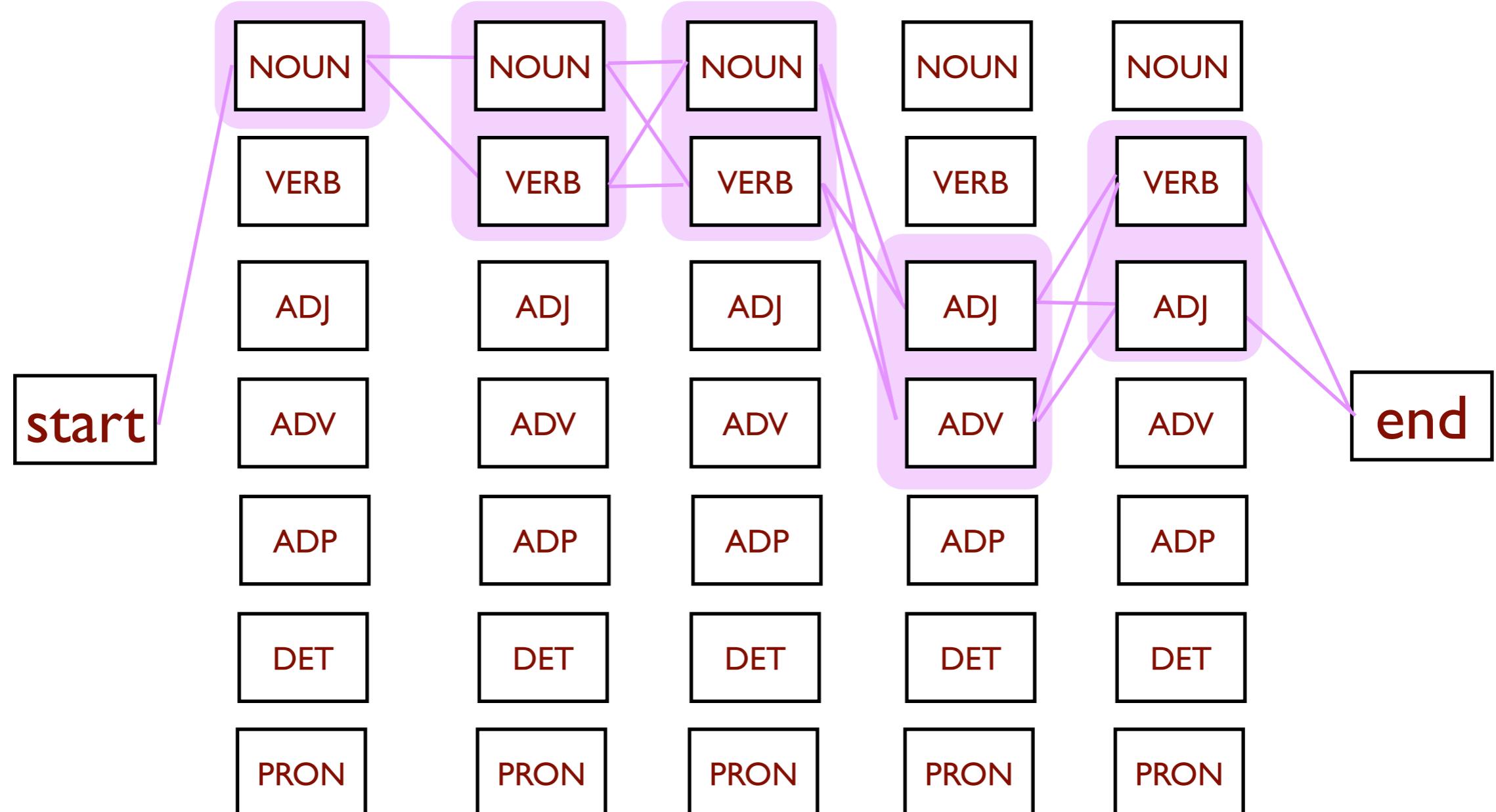


Coupled Token and Type Constraints



Produkterna
måste vara
helt
rena

ADJ VERB VERB ADJ

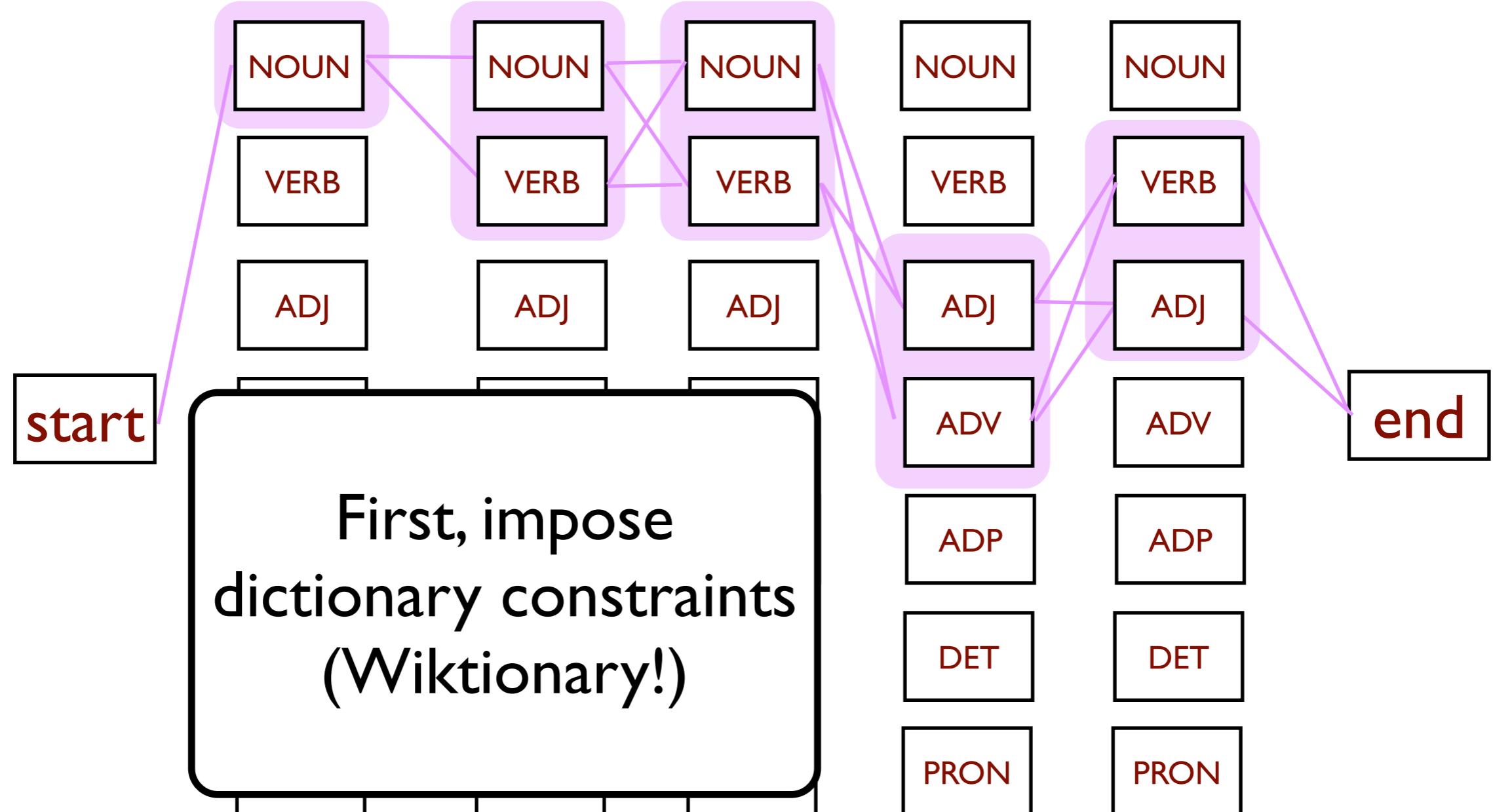


Coupled Token and Type Constraints

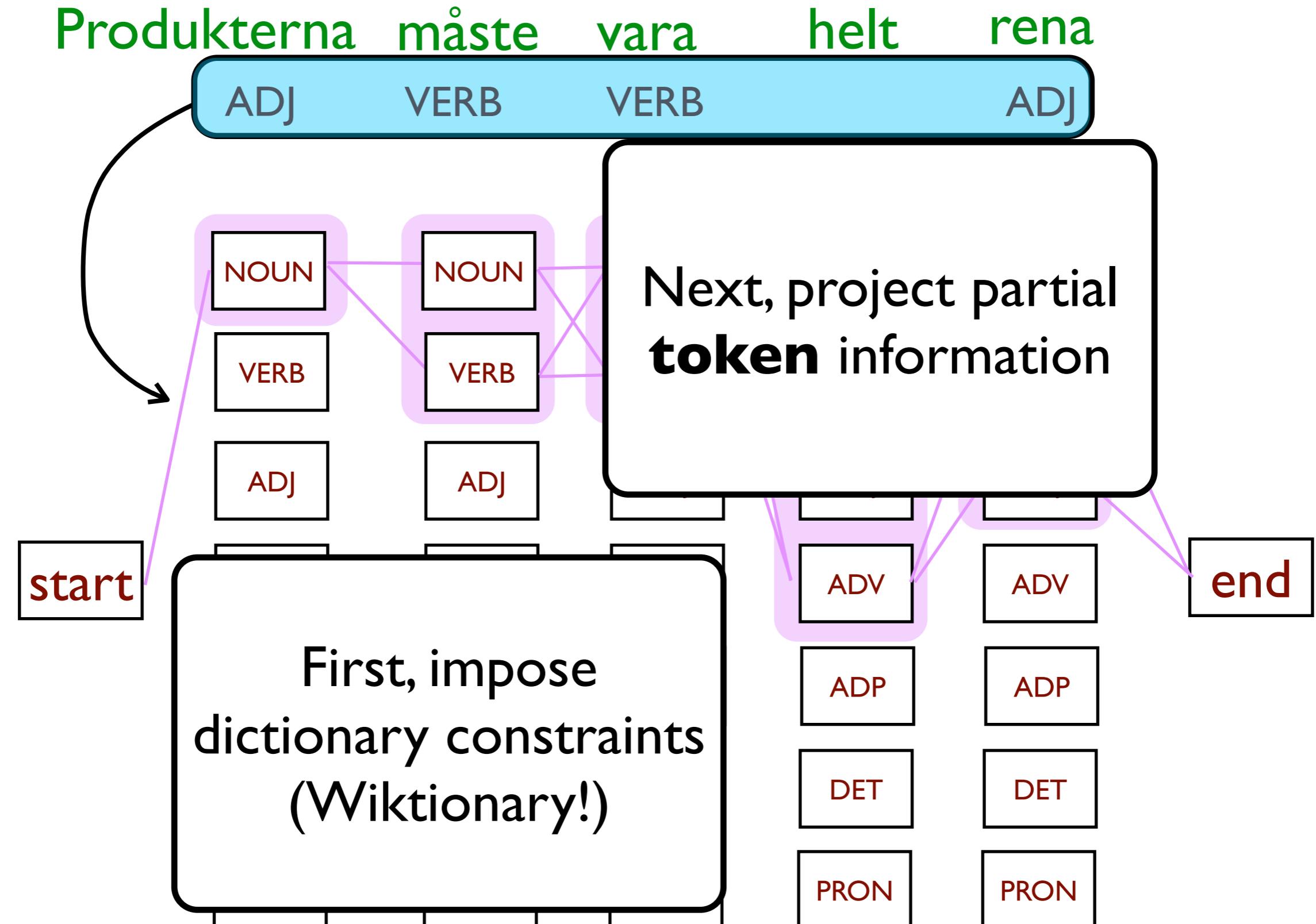


Produkterna
måste vara
helt
rena

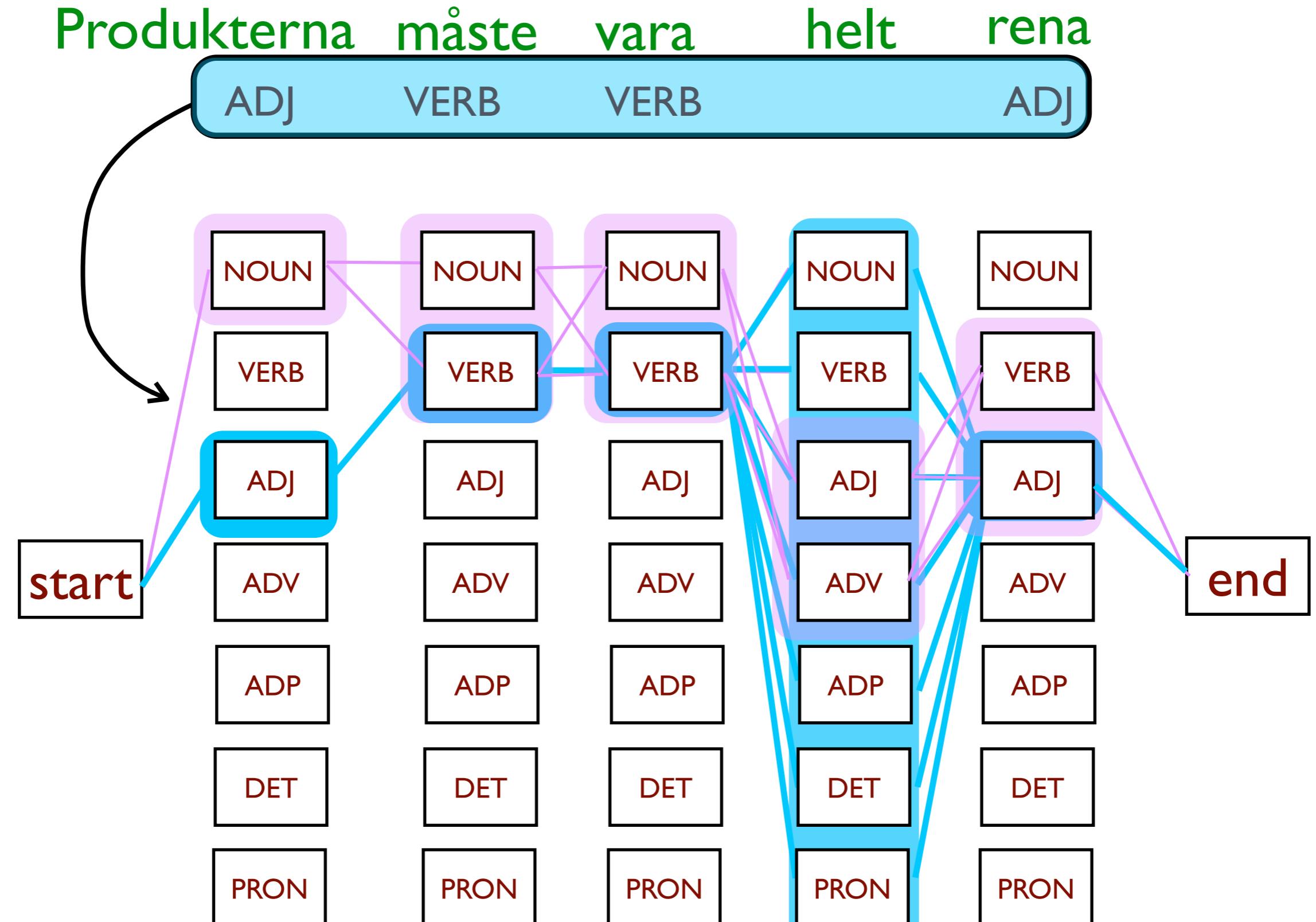
ADJ VERB VERB ADJ



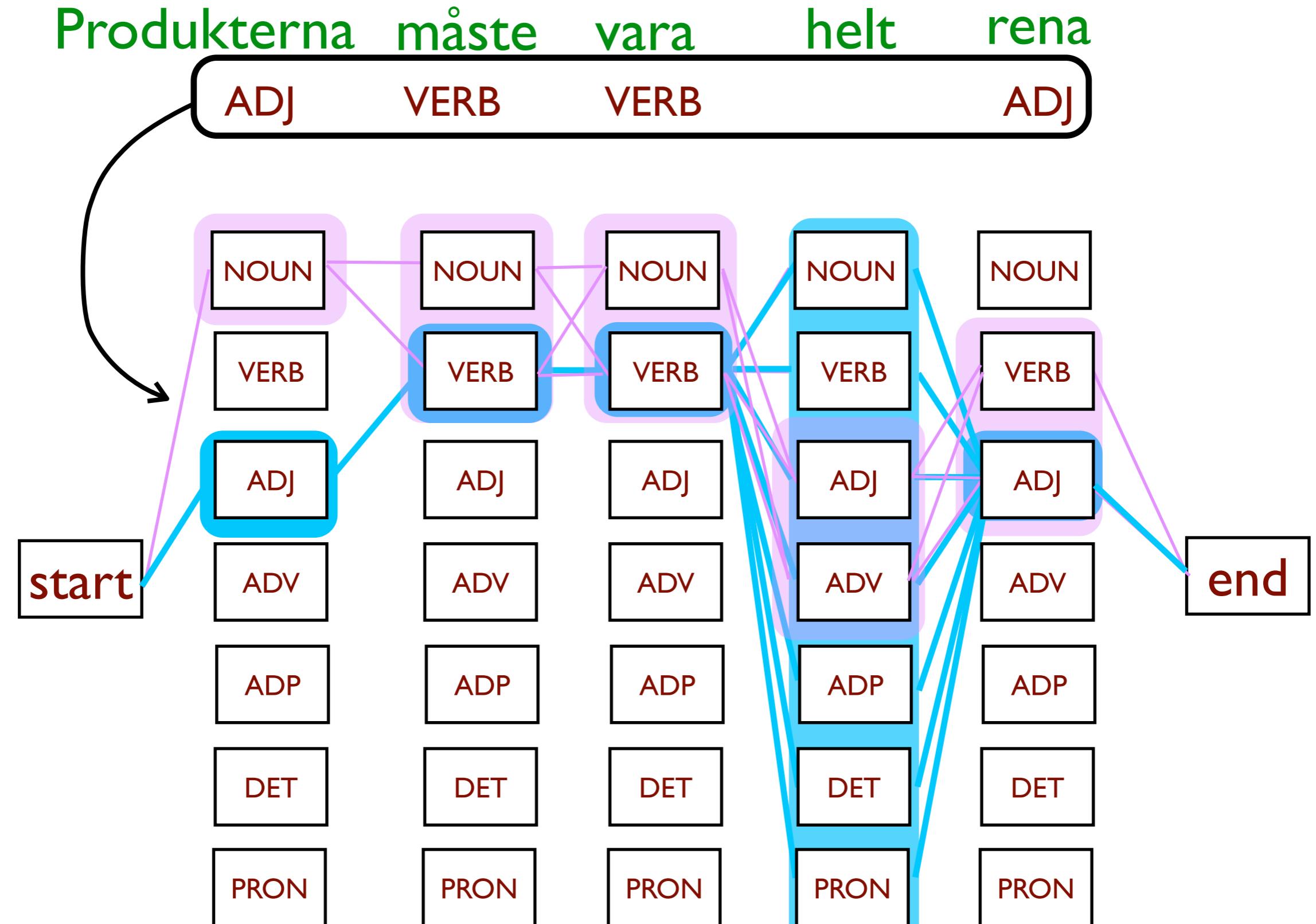
Coupled Token and Type Constraints



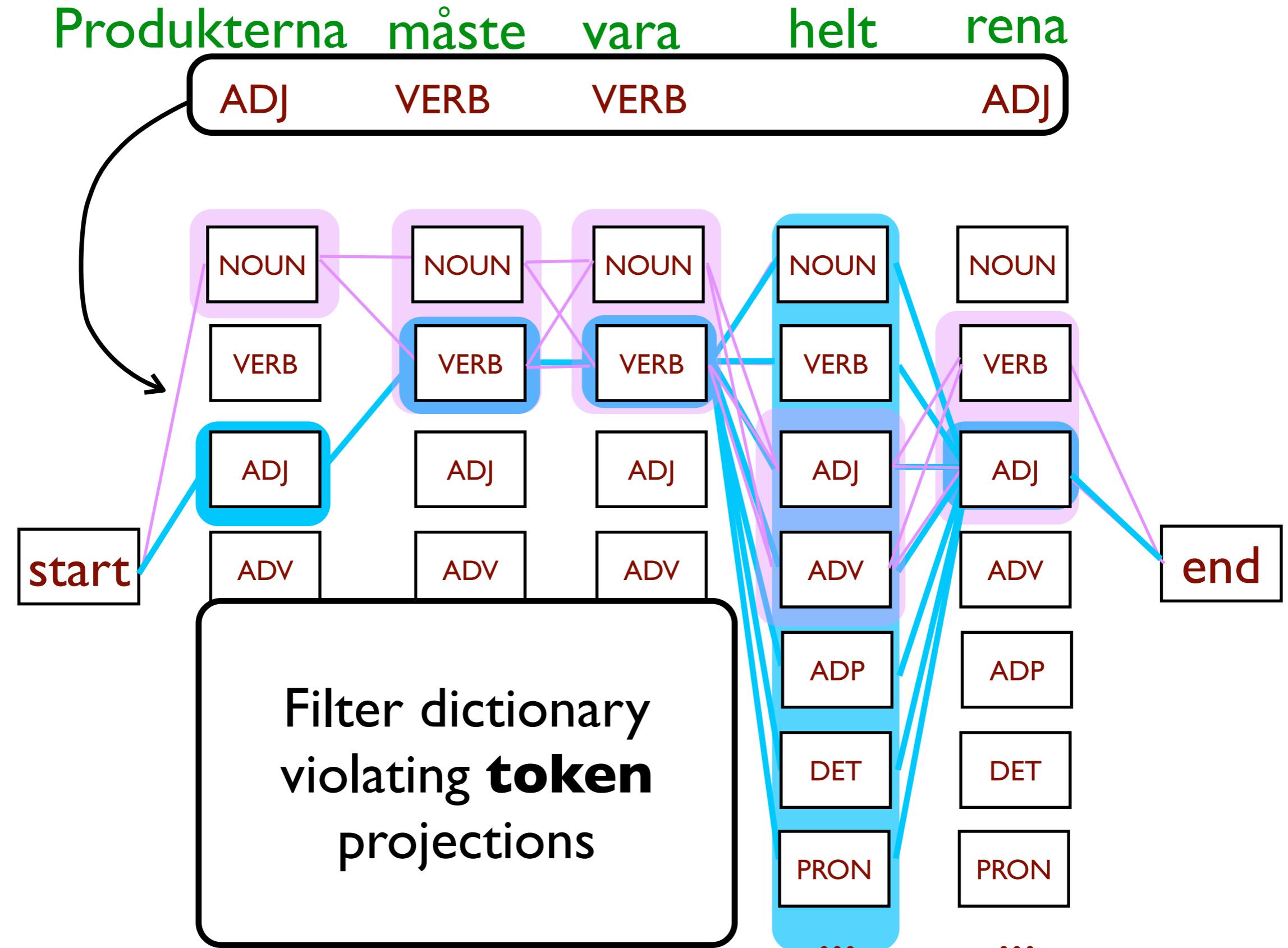
Coupled Token and Type Constraints



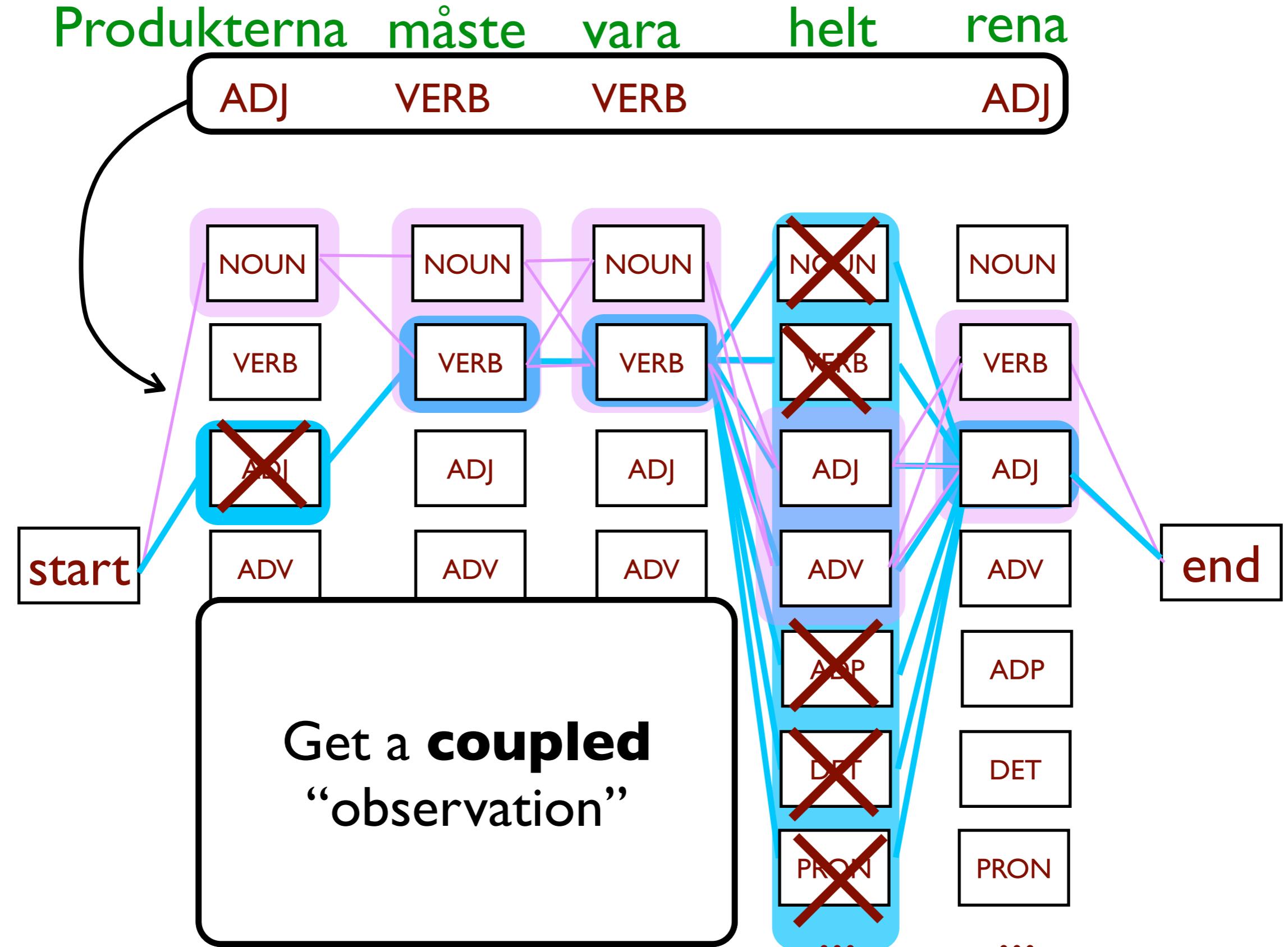
Coupled Token and Type Constraints



Coupled Token and Type Constraints

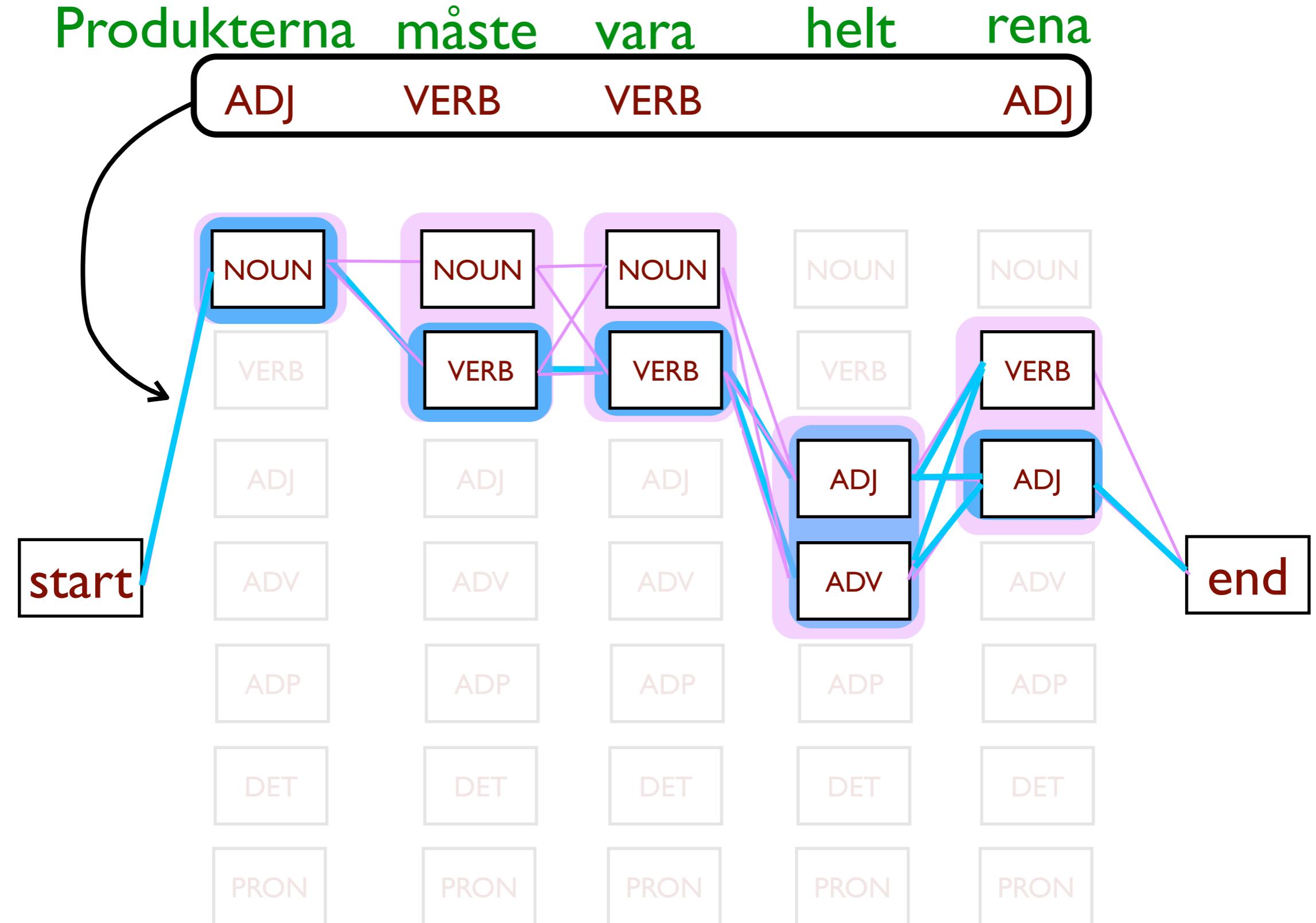


Coupled Token and Type Constraints



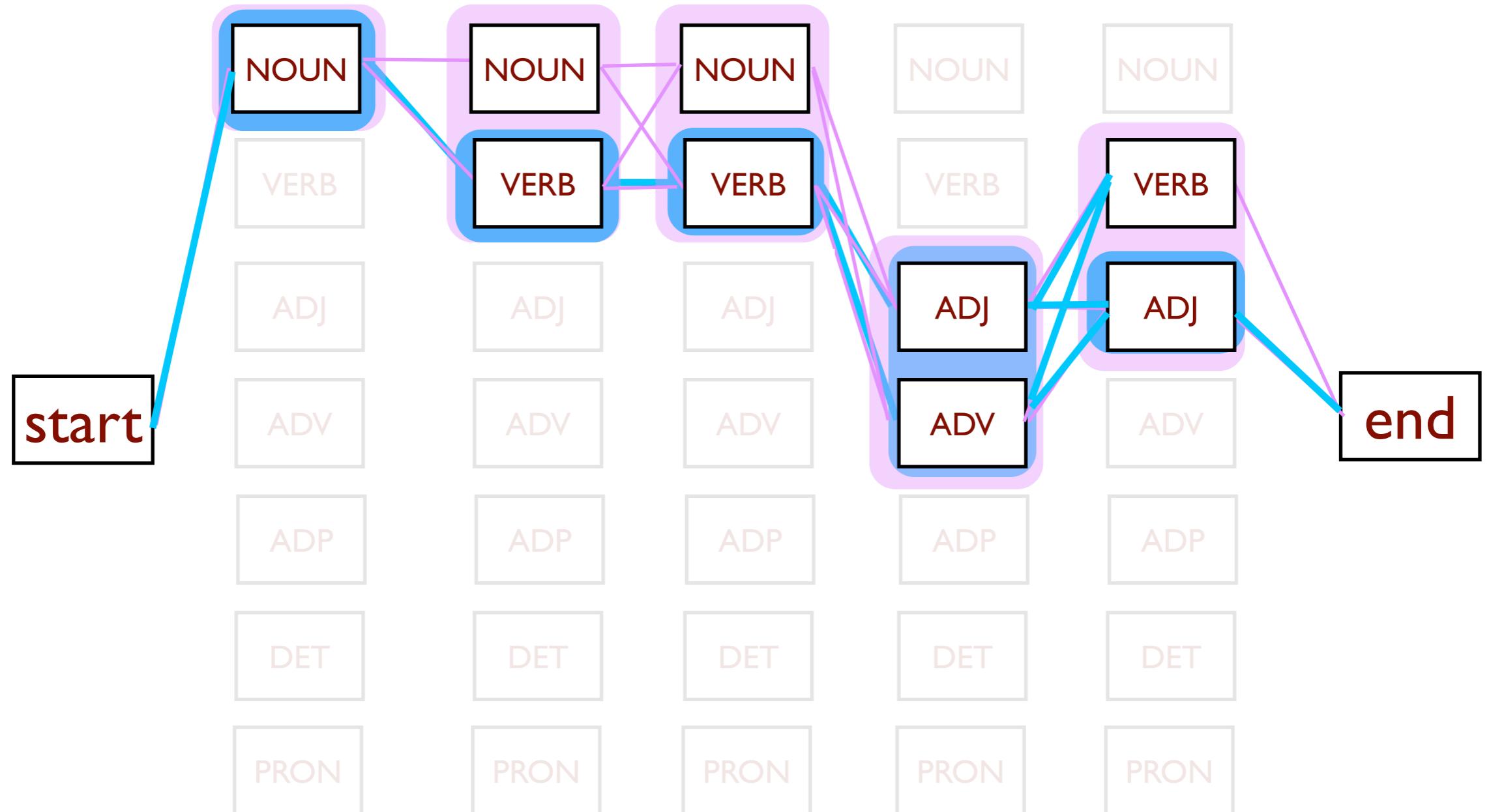
Coupled Token and Type Constraints

Google™



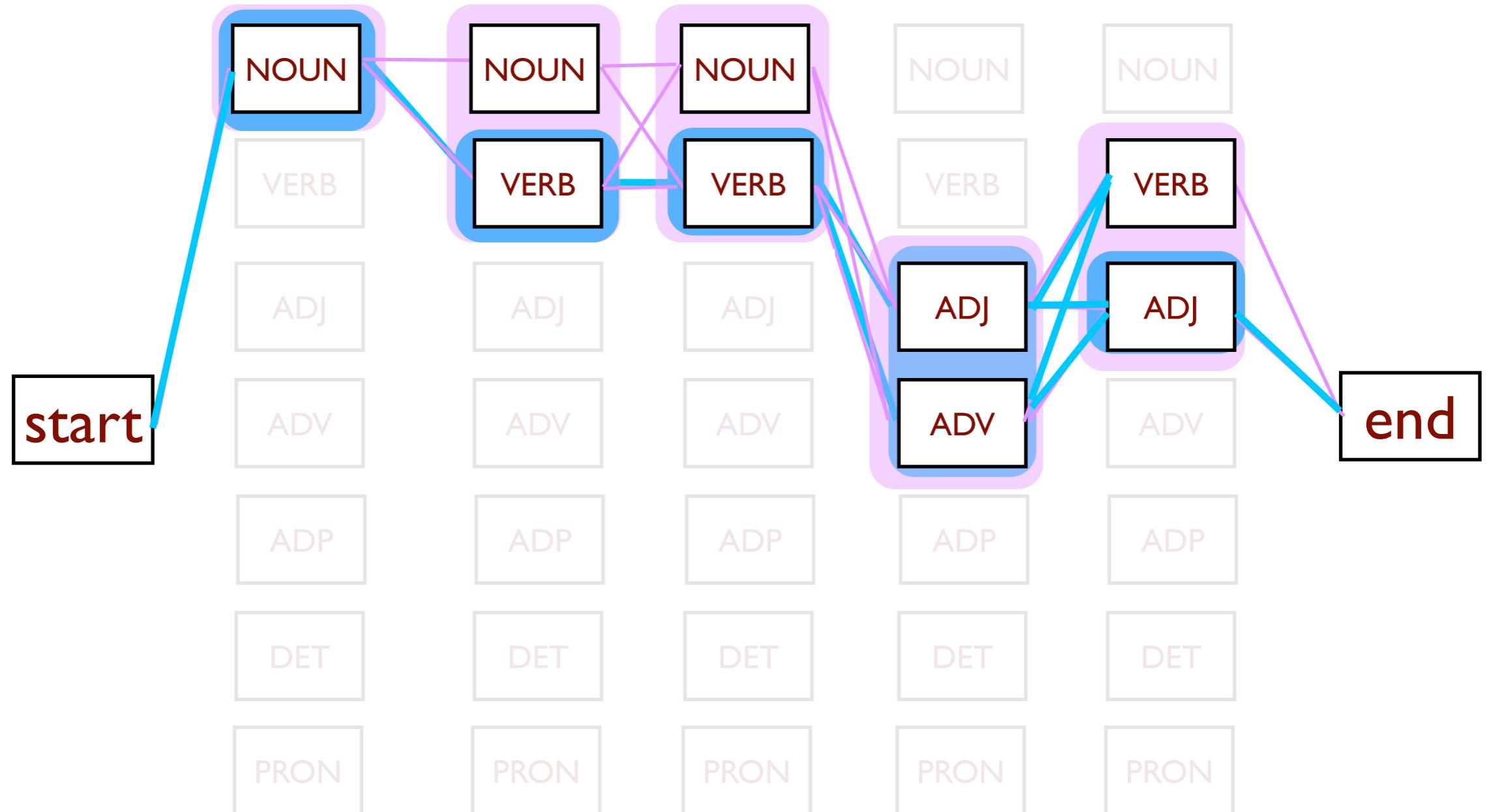
Coupled Token and Type Constraints

Google™



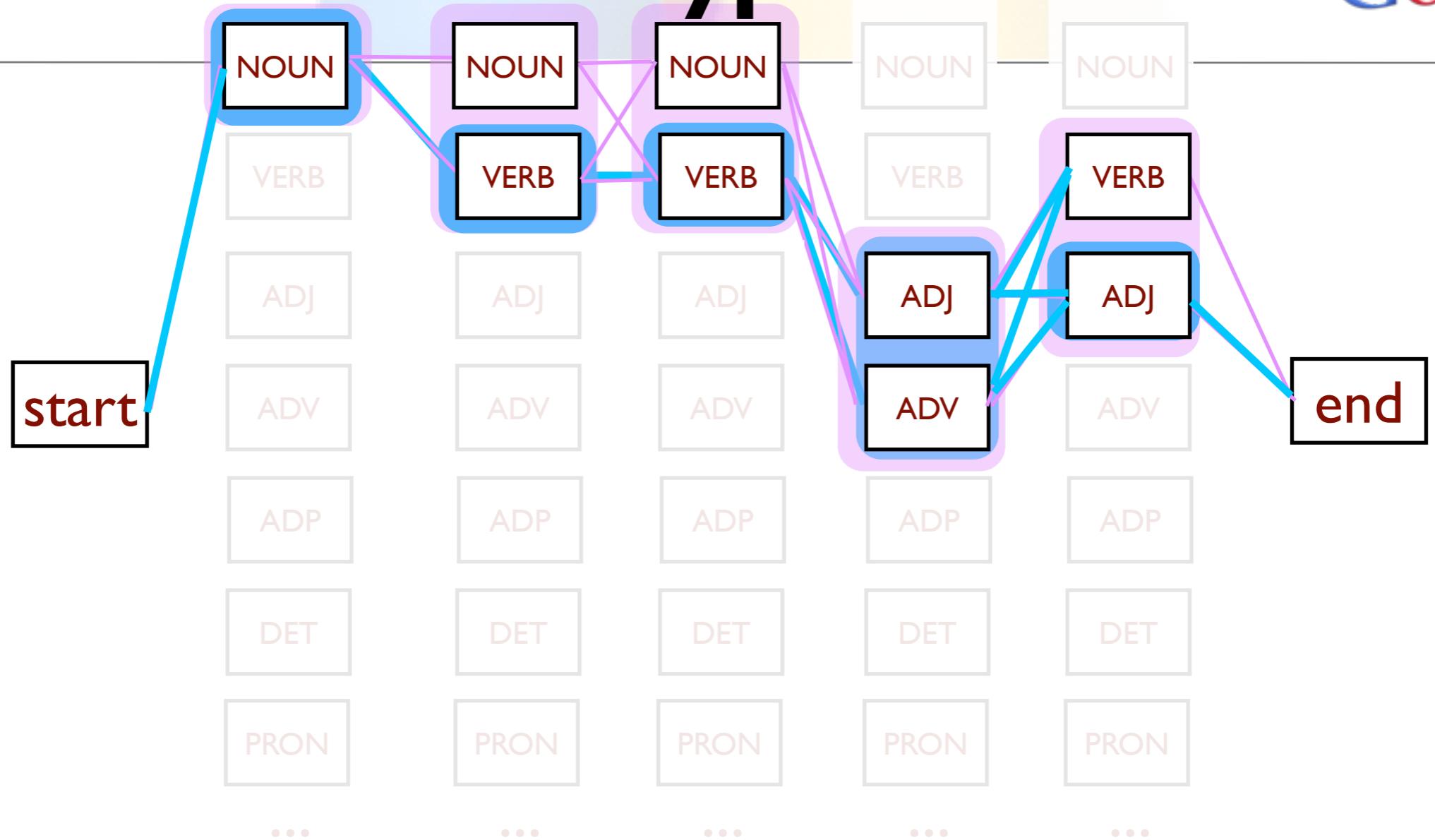
Coupled Token and Type Constraints

Google™



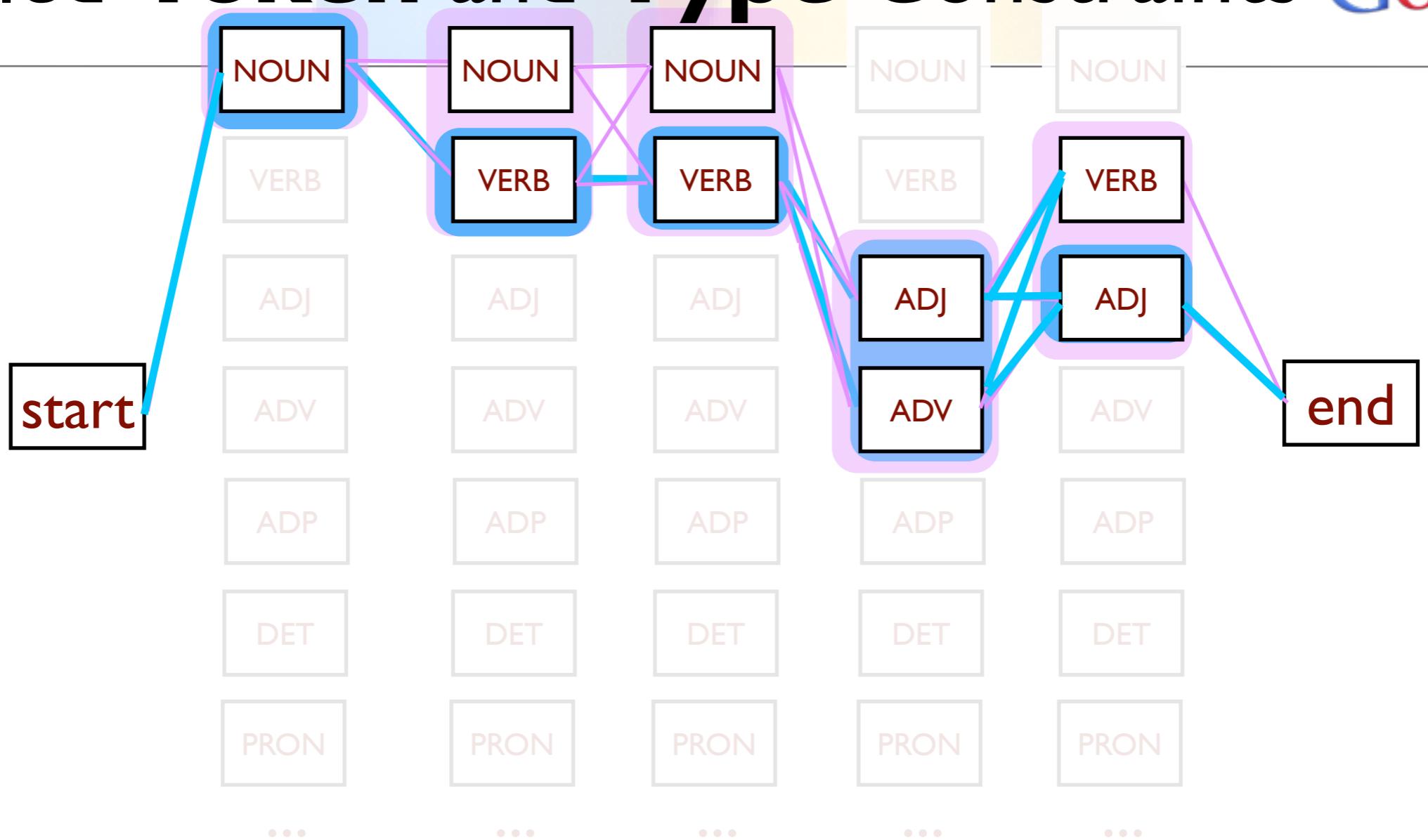
Coupled Token and Type Constraints

Google™



Coupled Token and Type Constraints

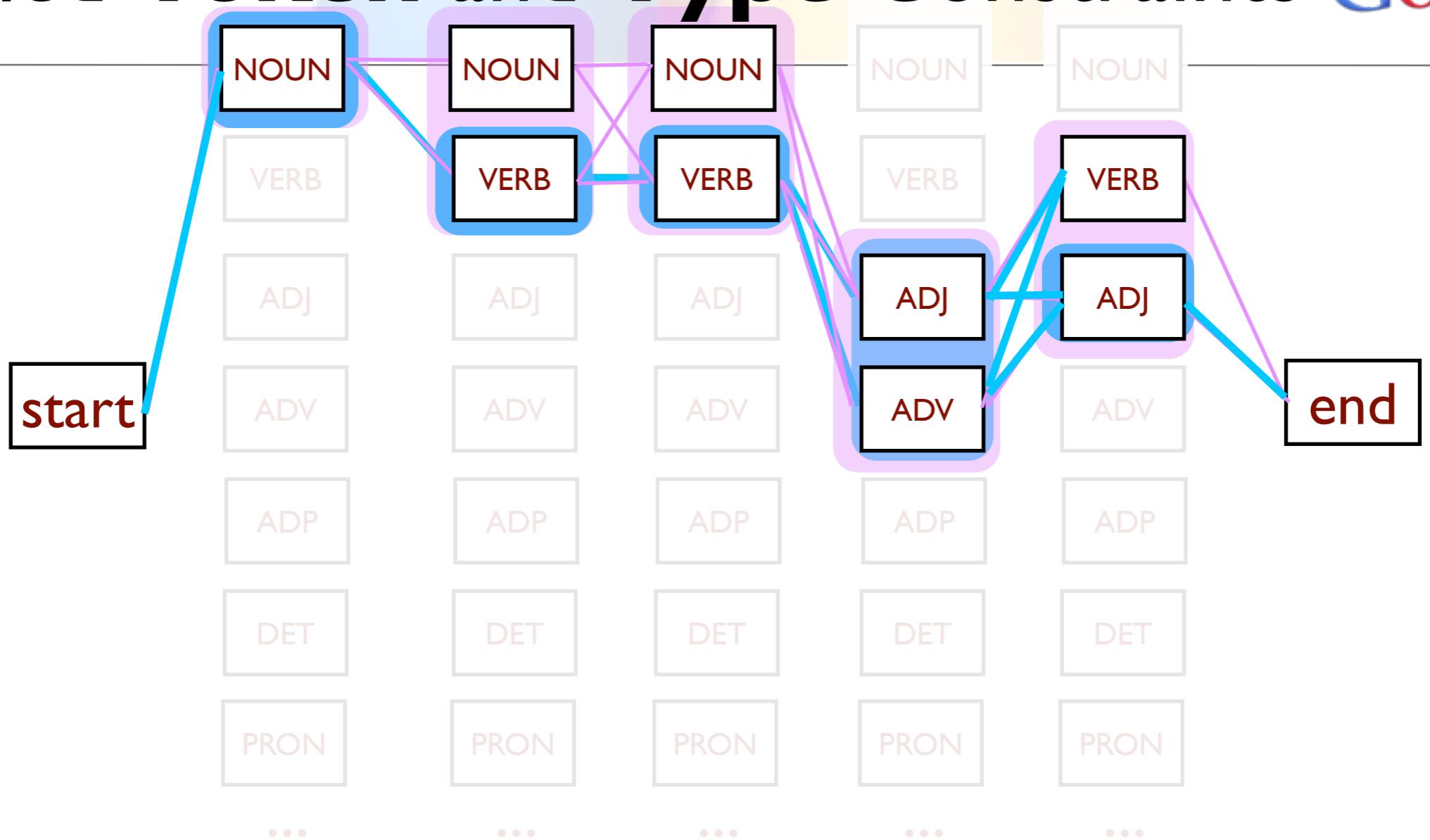
Google™



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

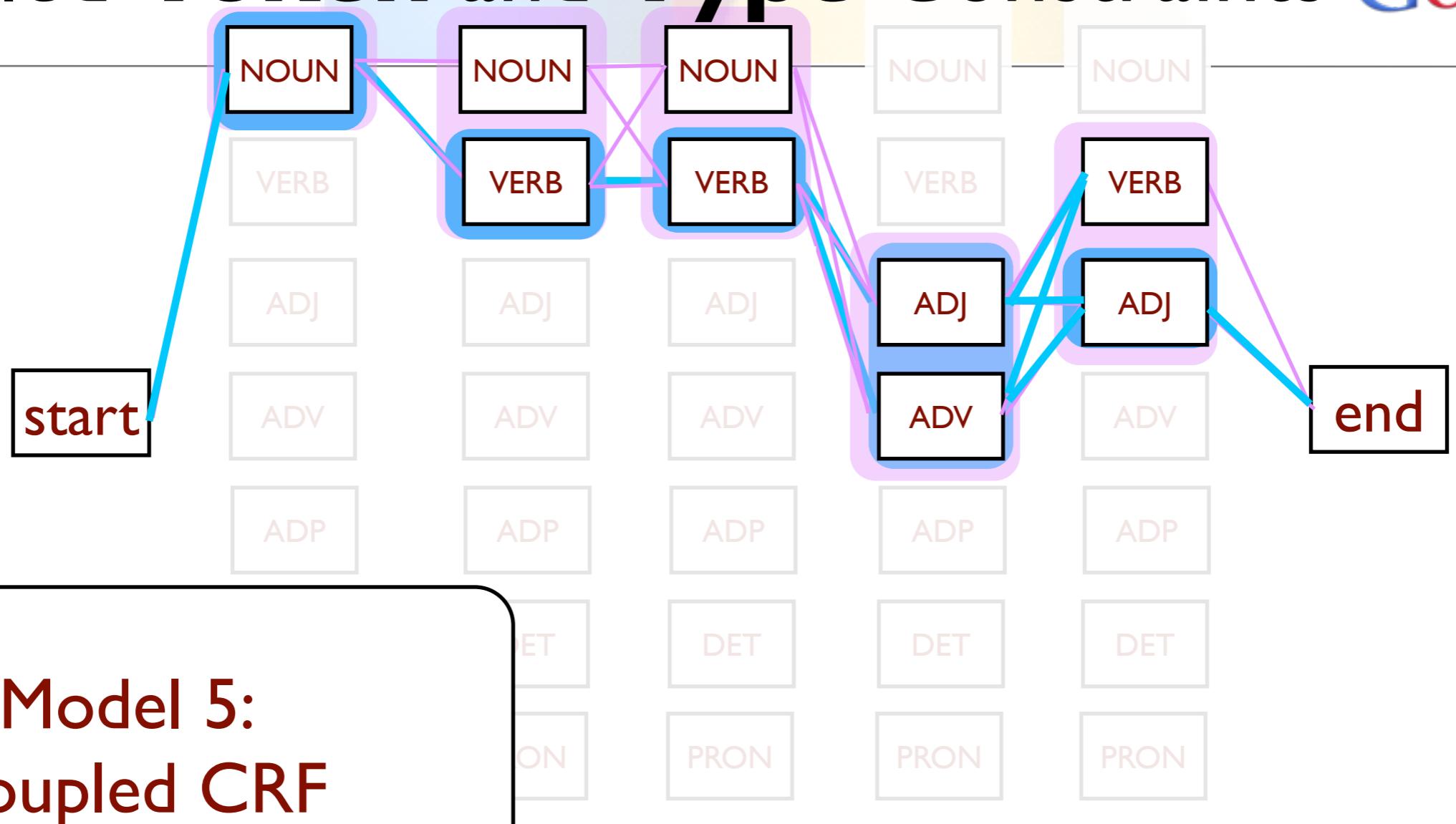
Coupled Token and Type Constraints

Google™



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

Coupled Token and Type Constraints



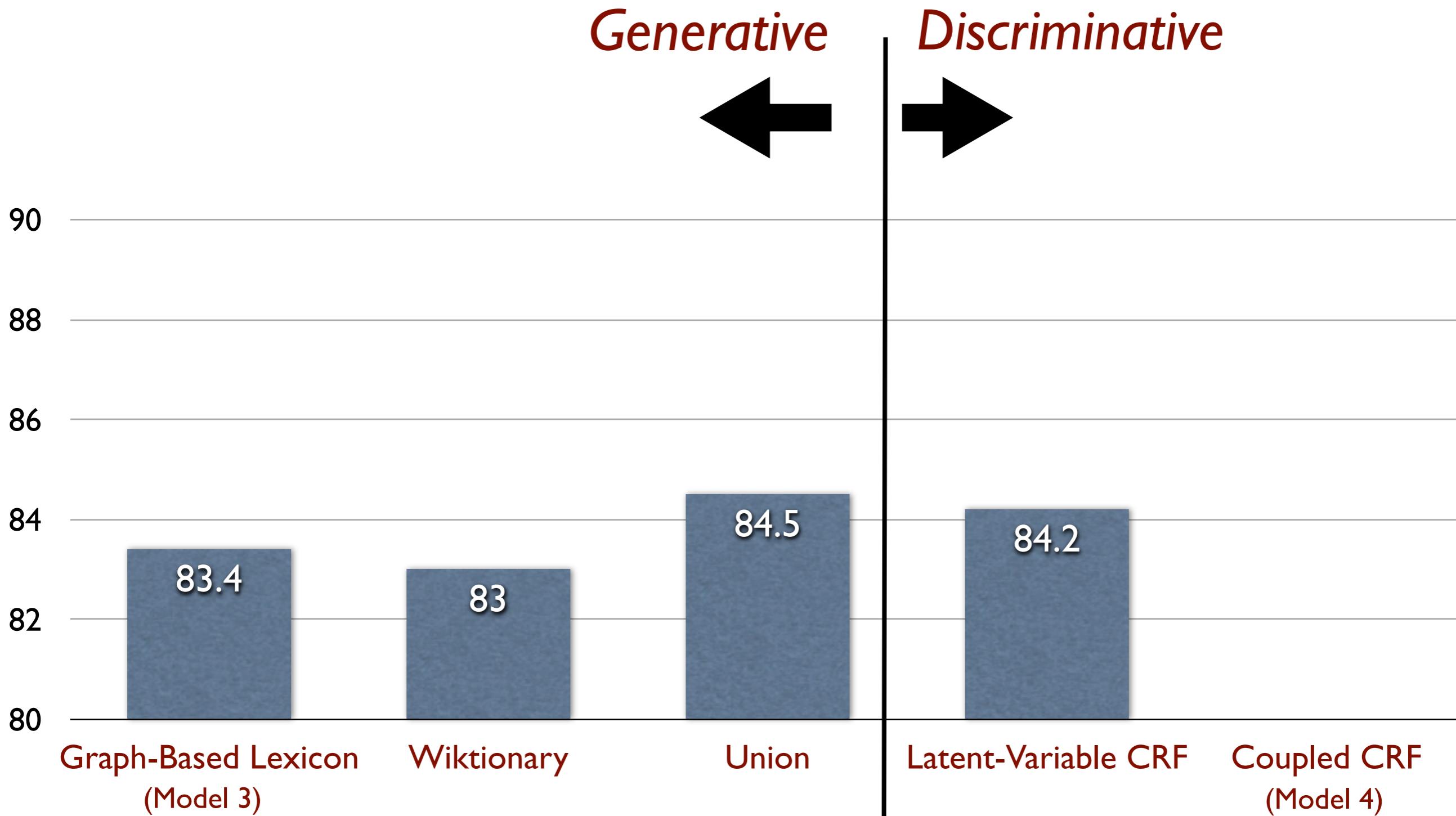
Model 5:
Coupled CRF

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

Model Comparison



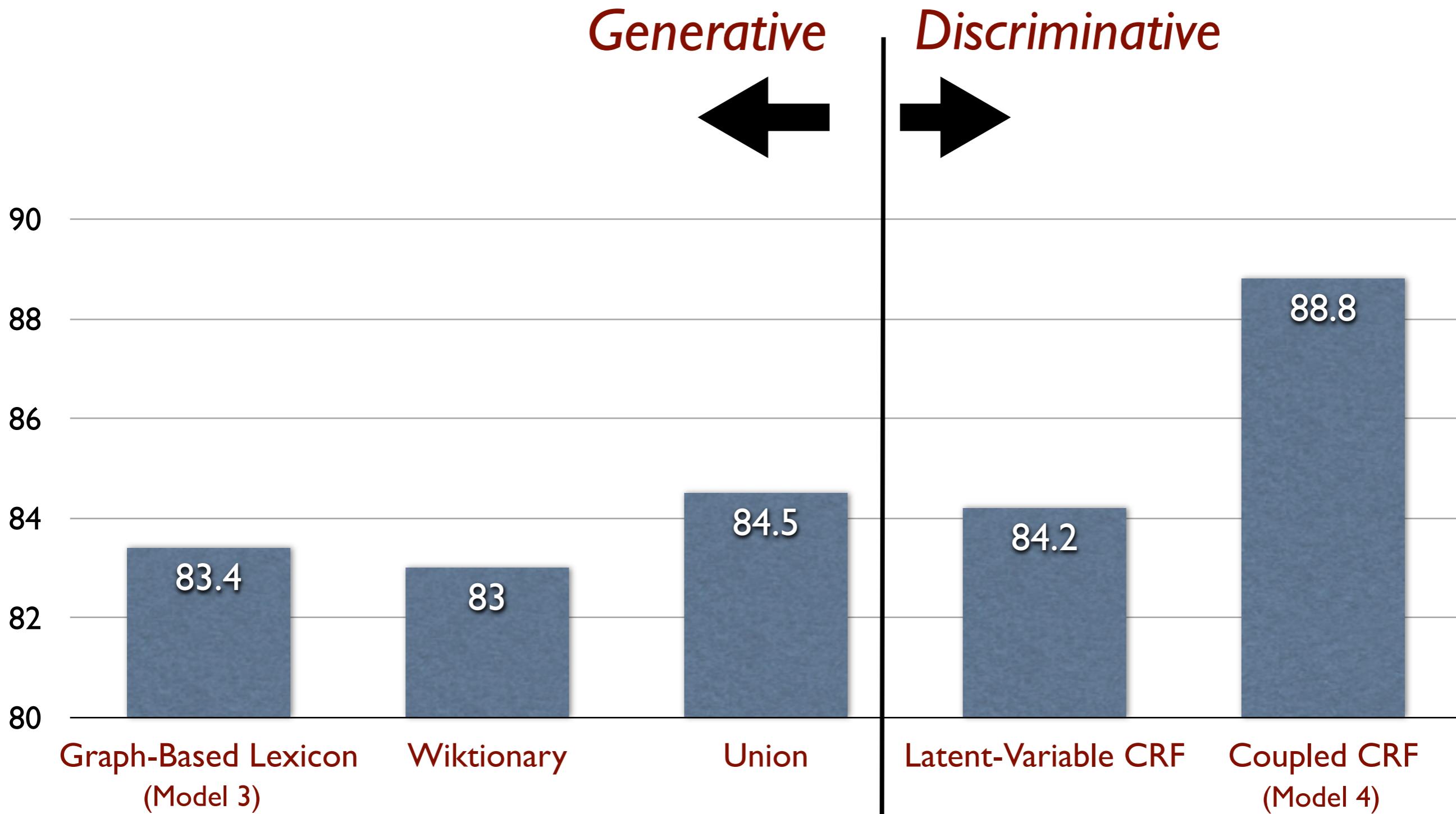
Measured across 8 Indo-European Languages



Model Comparison



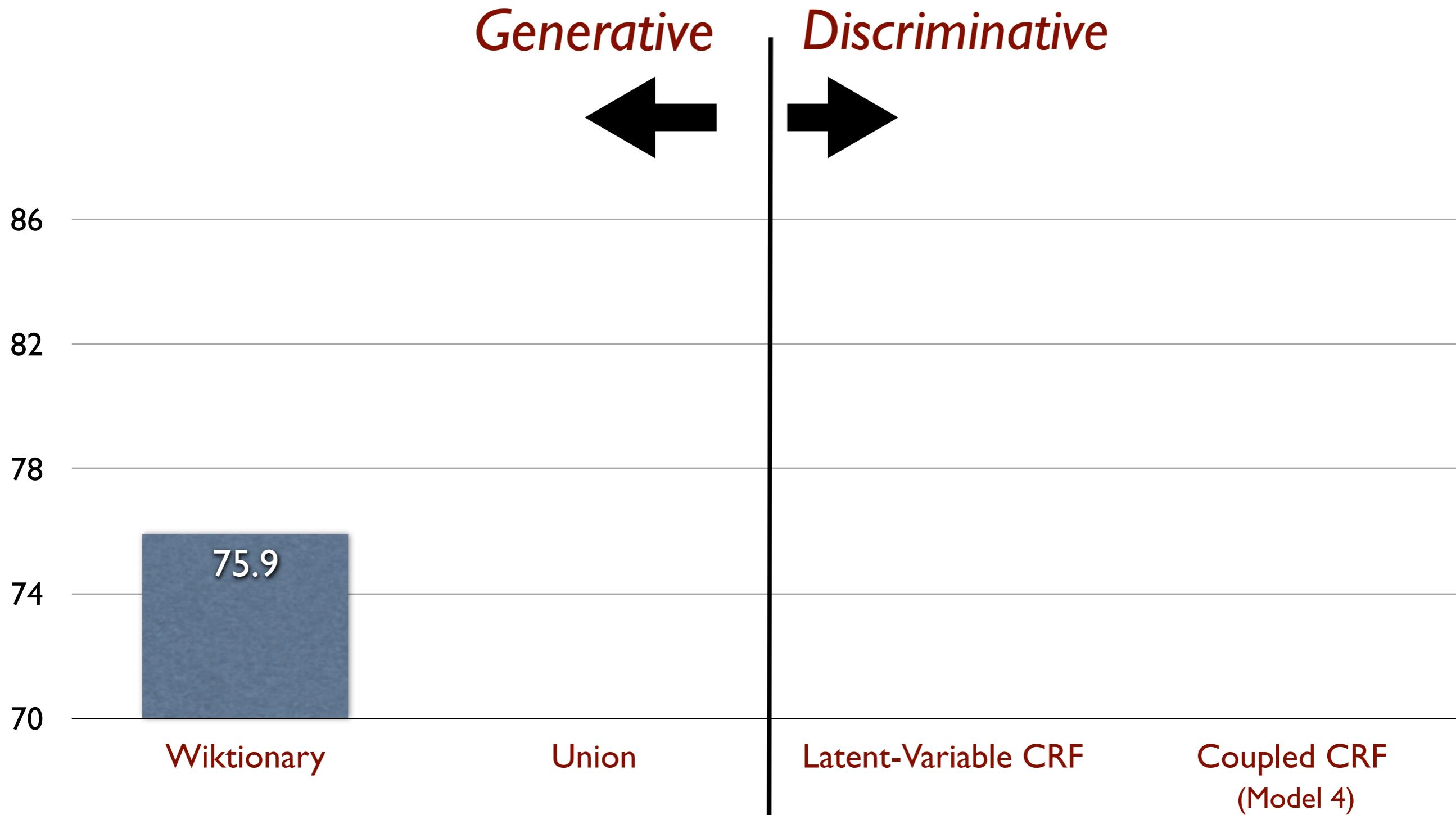
Measured across 8 Indo-European Languages



Model Comparison

Google™

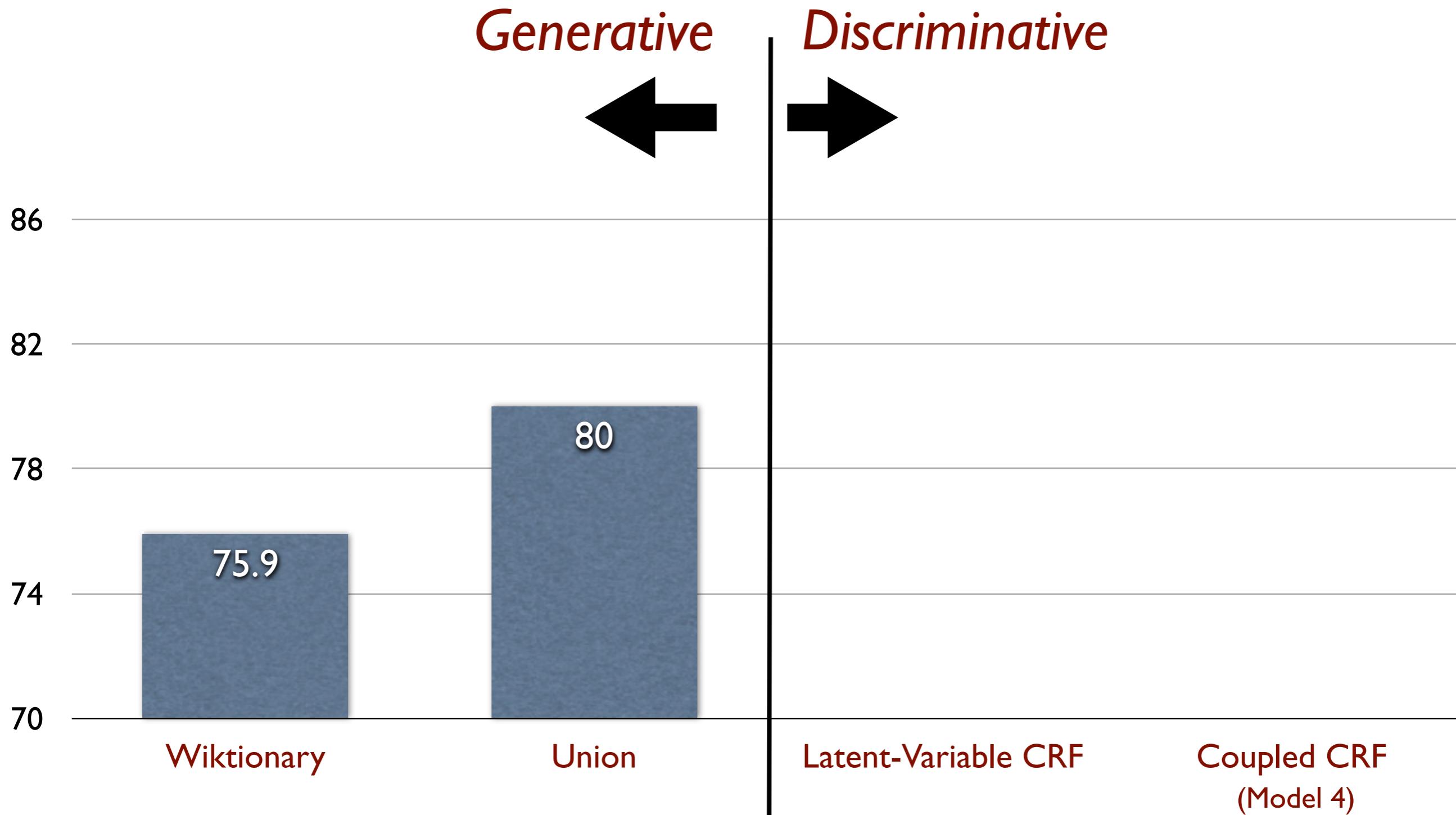
Measured across 15 Languages



Model Comparison

Google™

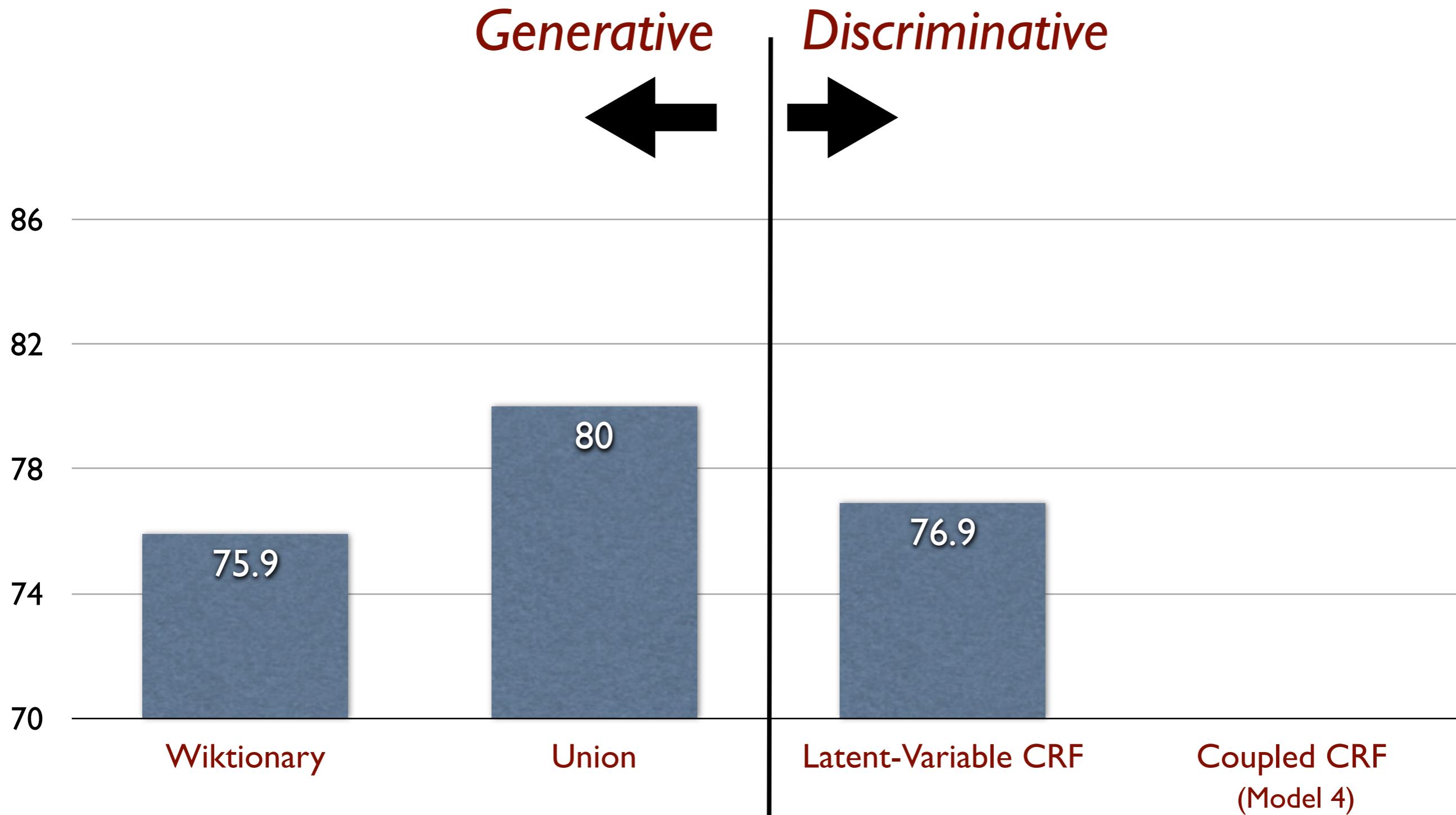
Measured across 15 Languages



Model Comparison

Google™

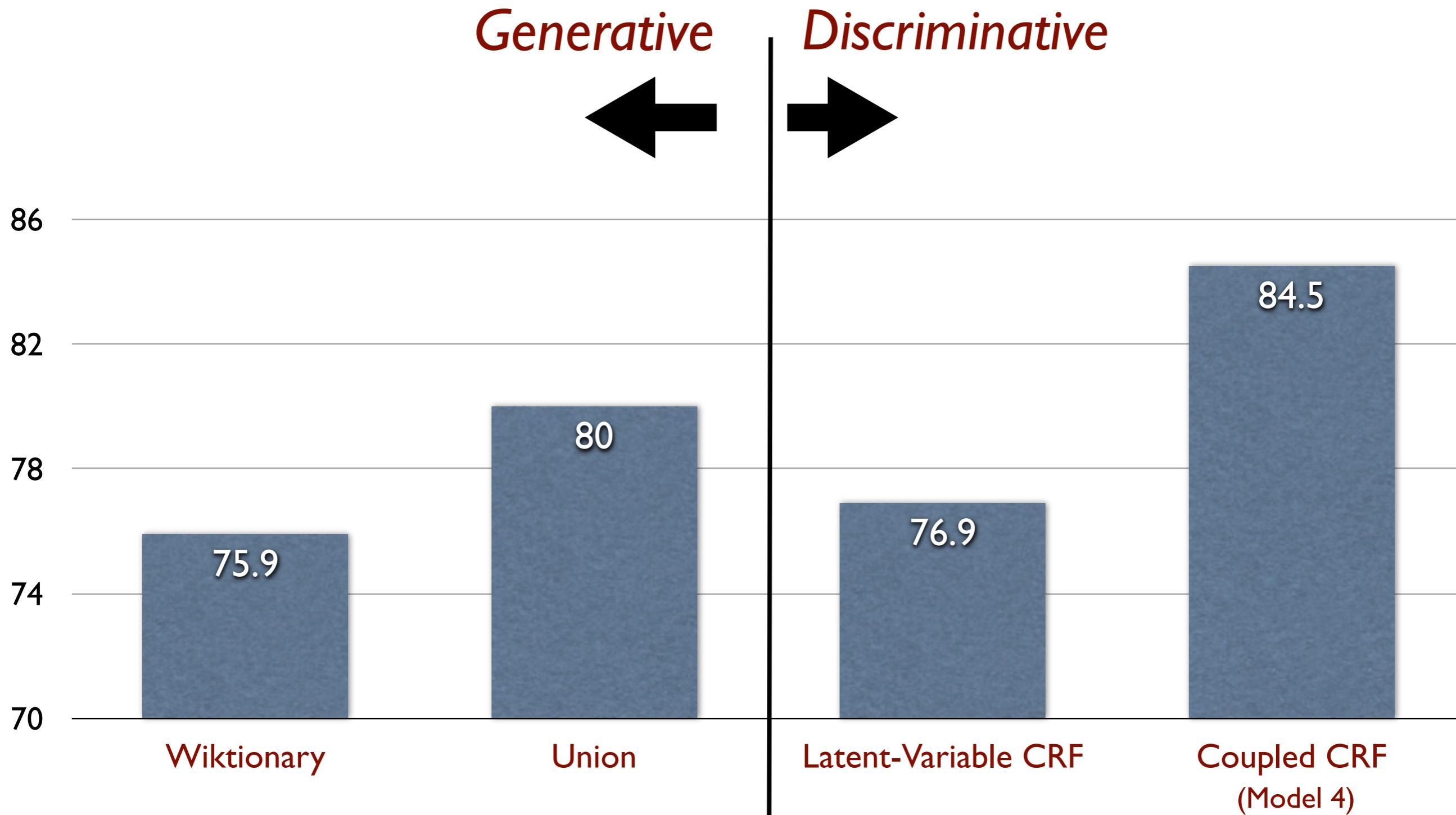
Measured across 15 Languages



Model Comparison

Google™

Measured across 15 Languages



Learning Part-of-Speech Taggers with Projected Dictionaries

Incorporating Hard Constraints from Translated sentences
and Dictionaries

Generalizing above with *Posterior Regularization*

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) :

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

Posterior Regularization



In a discriminative model (CRF) :

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

Entire dataset:

$$(X, Y) = ((x_1, y_1) \dots (x_n, y_n))$$

Posterior Regularization



In a discriminative model (CRF) :

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

Entire dataset:

$$(\mathbf{X}, \mathbf{Y}) = ((x_1, y_1) \dots (x_n, y_n))$$

Maximum conditional likelihood:

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

Posterior Regularization



Supervised Objective:

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

Posterior Regularization



Supervised Objective:

$$\begin{aligned}\max_{\theta} \mathcal{L}(\theta) &= \max_{\theta} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma \|\theta\| \\ &= \min_{\theta} \mathcal{D}_{\text{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma \|\theta\| + \text{constant}\end{aligned}$$

Posterior Regularization



Supervised Objective:

$$\begin{aligned}\max_{\theta} \mathcal{L}(\theta) &= \max_{\theta} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma \|\theta\| \\ &= \min_{\theta} \mathcal{D}_{\text{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma \|\theta\| + \text{constant}\end{aligned}$$

Distribution with
probability of labeled
data = 1

Posterior Regularization



Supervised Objective:

$$\begin{aligned}\max_{\theta} \mathcal{L}(\theta) &= \max_{\theta} \log(p(\mathbf{Y}|\mathbf{X})) - \gamma \|\theta\| \\ &= \min_{\theta} \mathcal{D}_{\text{KL}}(q|p(\mathbf{Y}|\mathbf{X})) + \gamma \|\theta\| + \text{constant}\end{aligned}$$

Distribution with
probability of labeled
data = 1

What if we only have partial labeling?

Posterior Regularization



Supervised Objective:

Distribution with probability of
labeled data = 1

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

Posterior Regularization



Supervised Objective:

Distribution with probability of
labeled data = 1

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

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

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

Objective with Partial Labelings:

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

$$\begin{aligned} & \max_{\theta} \mathcal{L}(\theta) \\ &= \max_{\theta} \max_{q \in Q} -\mathcal{D}_{\text{KL}}(q \mid\mid p(\mathbf{Y} \mid \mathbf{X})) - \gamma \|\theta\| - \text{constant} \end{aligned}$$

Posterior Regularization



Example:

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

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

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

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

Posterior Regularization



Example:

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

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

a constraint
feature

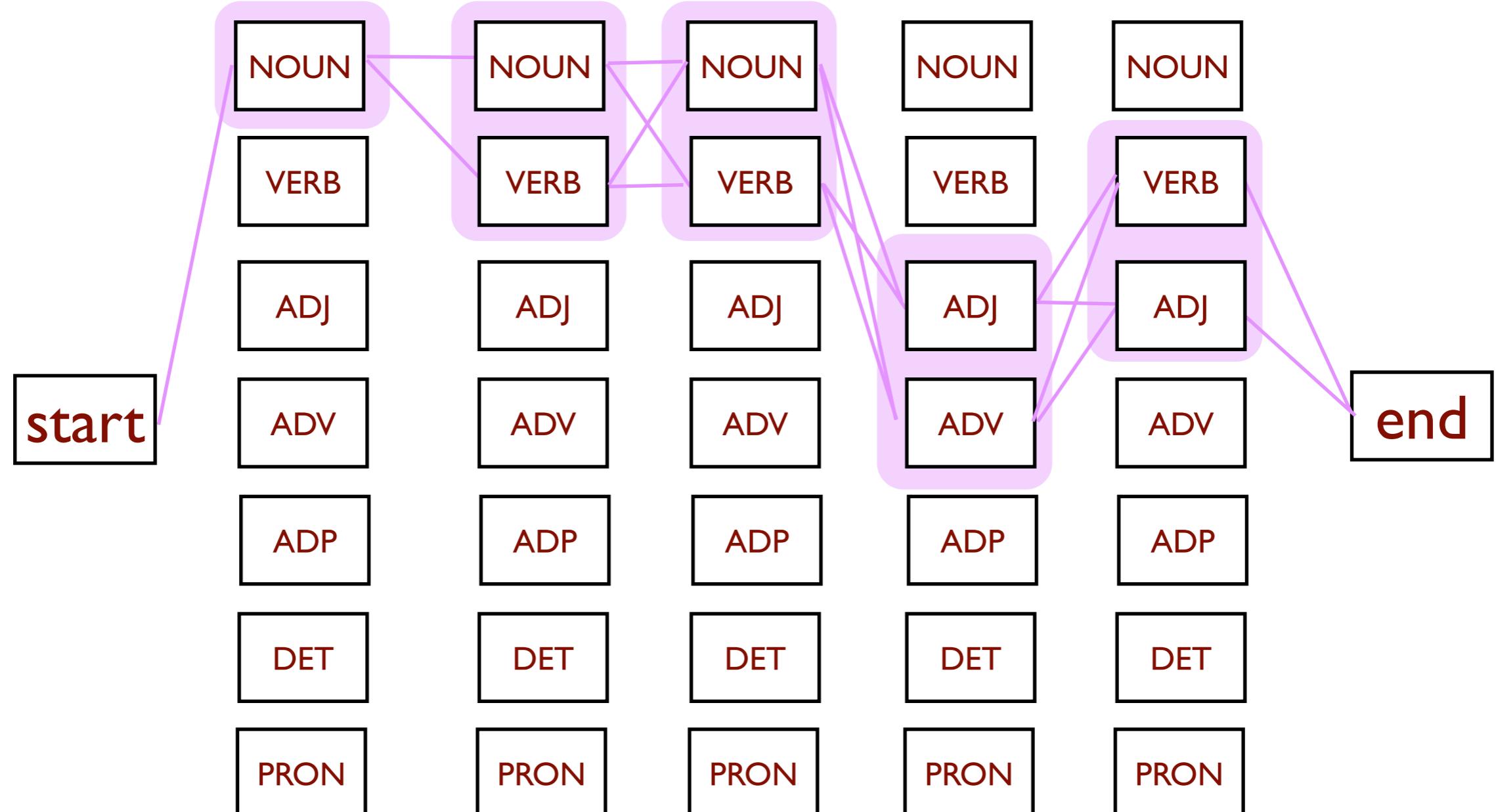
$$\mathcal{Q} = \{q(\mathbf{Y}) : \mathbf{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \geq 0.1\}$$

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

Posterior Regularization and Projection



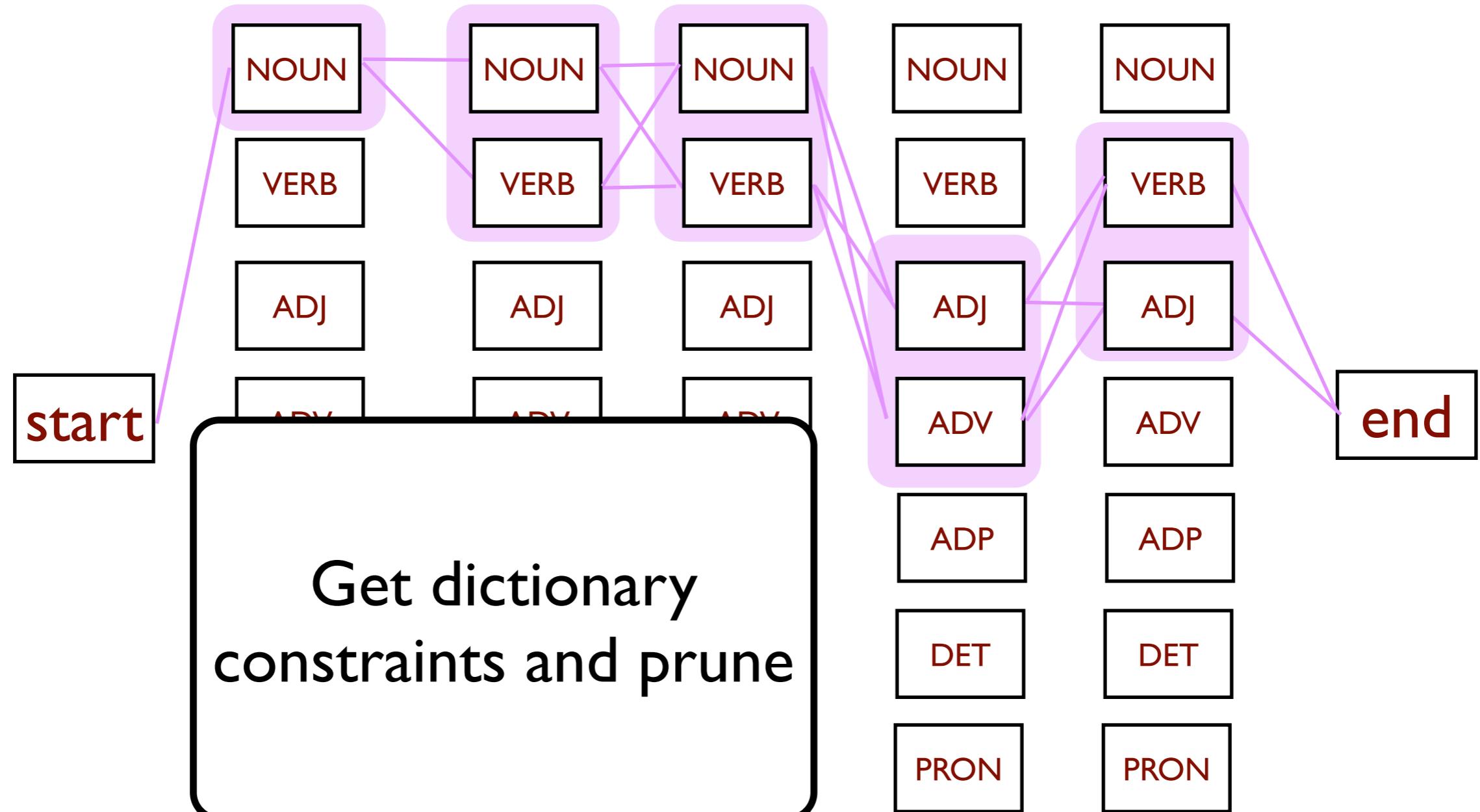
Produkterna måste vara helt rena
ADJ VERB VERB ADJ



Posterior Regularization and Projection



Produkterna måste vara helt rena
ADJ VERB VERB ADJ

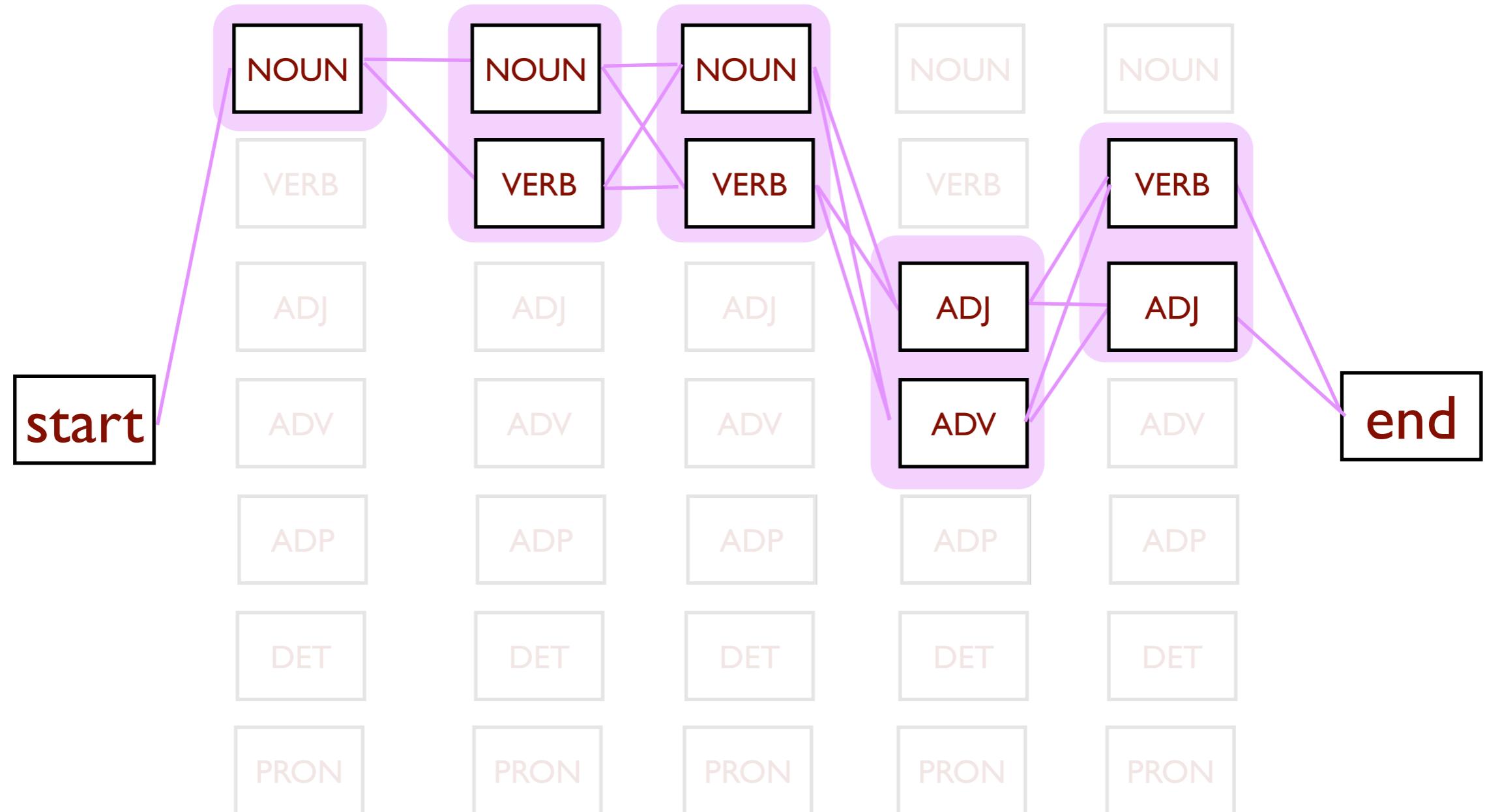


Posterior Regularization and Projection

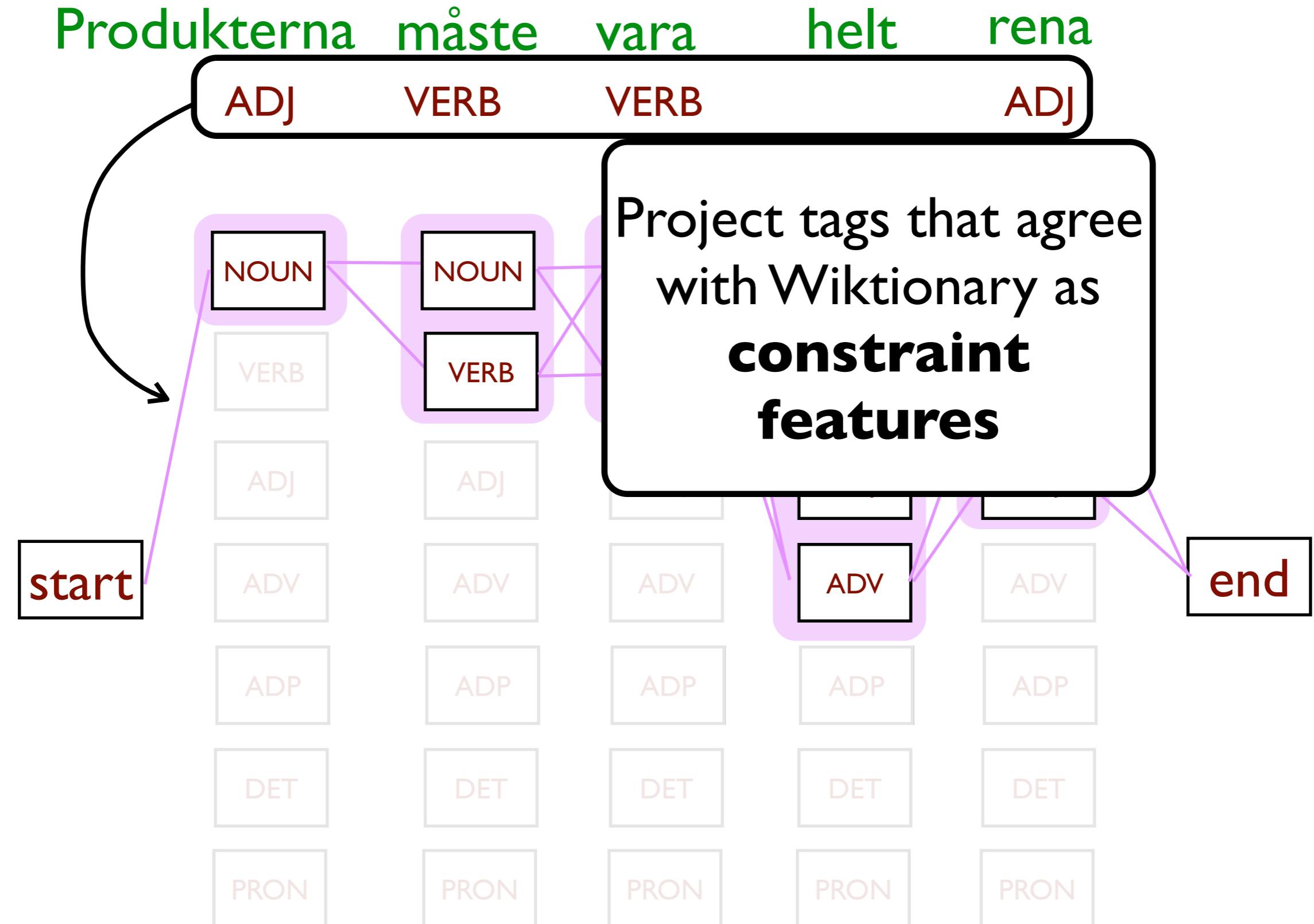


Produkterna
måste vara
helt
rena

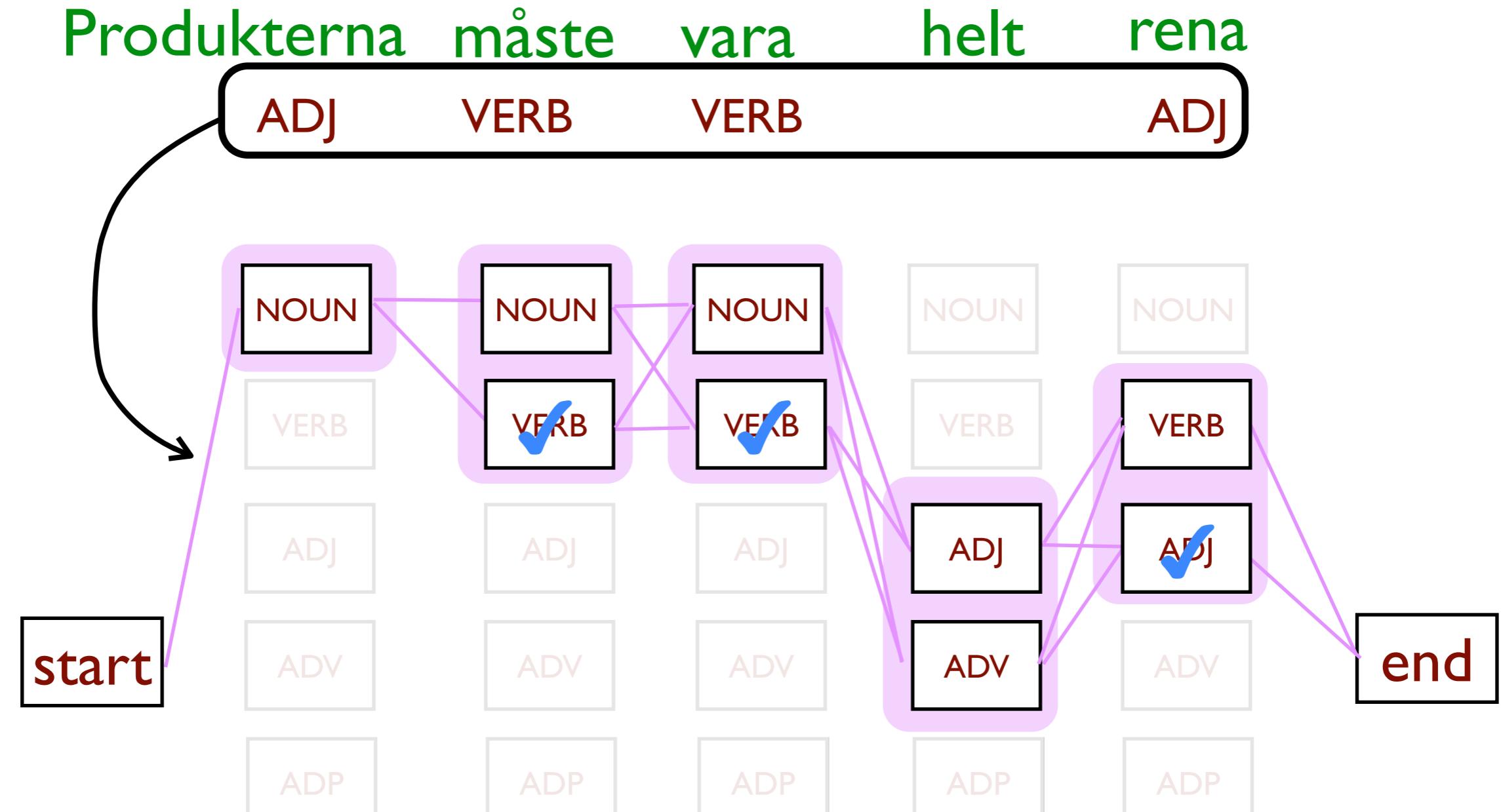
ADJ VERB VERB ADJ



Posterior Regularization and Projection



Posterior Regularization and Projection



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

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

Objective:

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

Objective:

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

Dual with respect to q :

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

Objective:

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

Dual with respect to q :

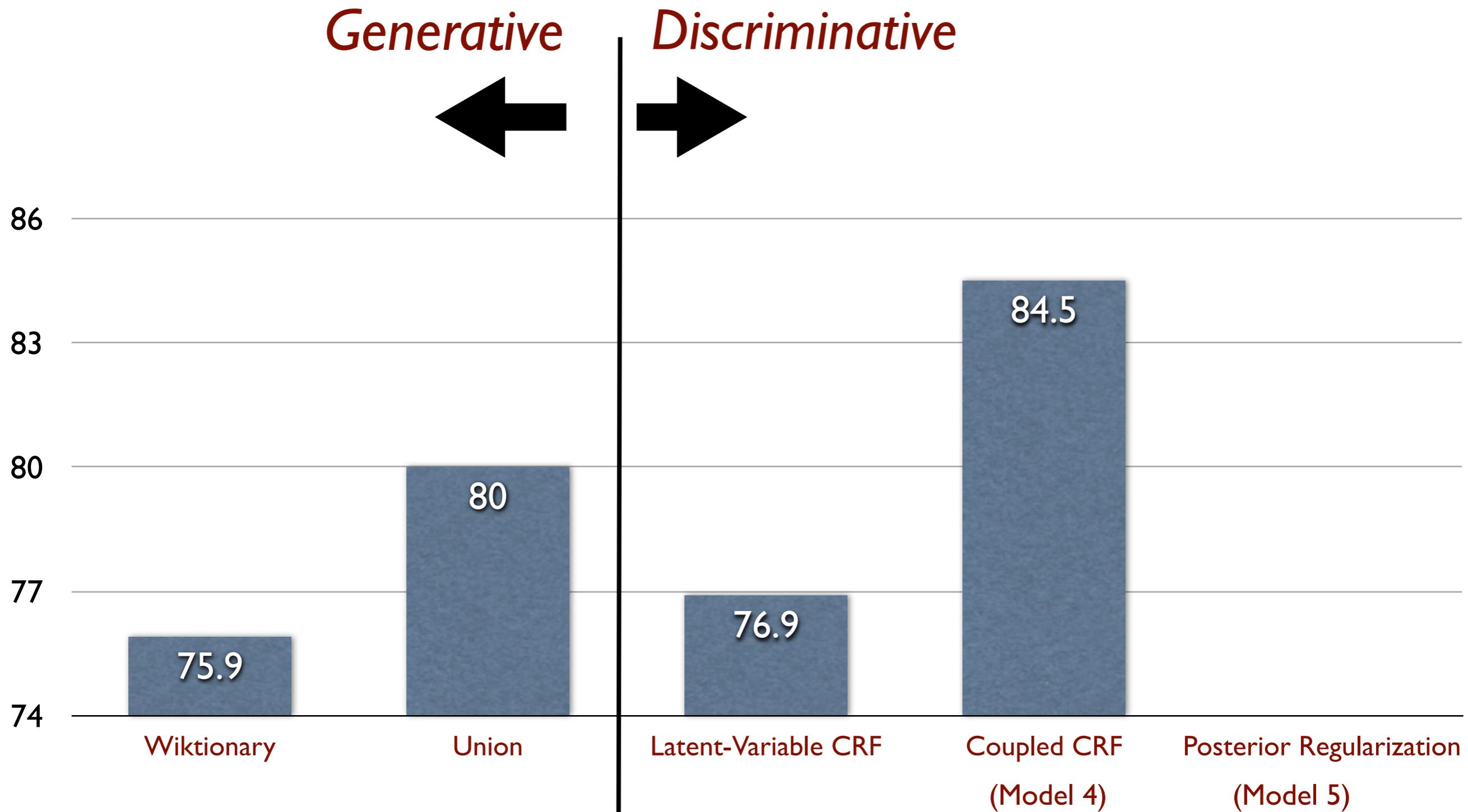
$$\max_{\theta} \min_{\lambda} \mathbf{b} \cdot \lambda + \sum_{\mathbf{Y}} p_{\theta}(\mathbf{Y} | \mathbf{X}) e^{-\lambda \cdot \phi(\mathbf{X}, \mathbf{Y})} - \gamma \|\theta\|$$

Optimized using alternating stochastic gradient descent

Model Comparison

Google™

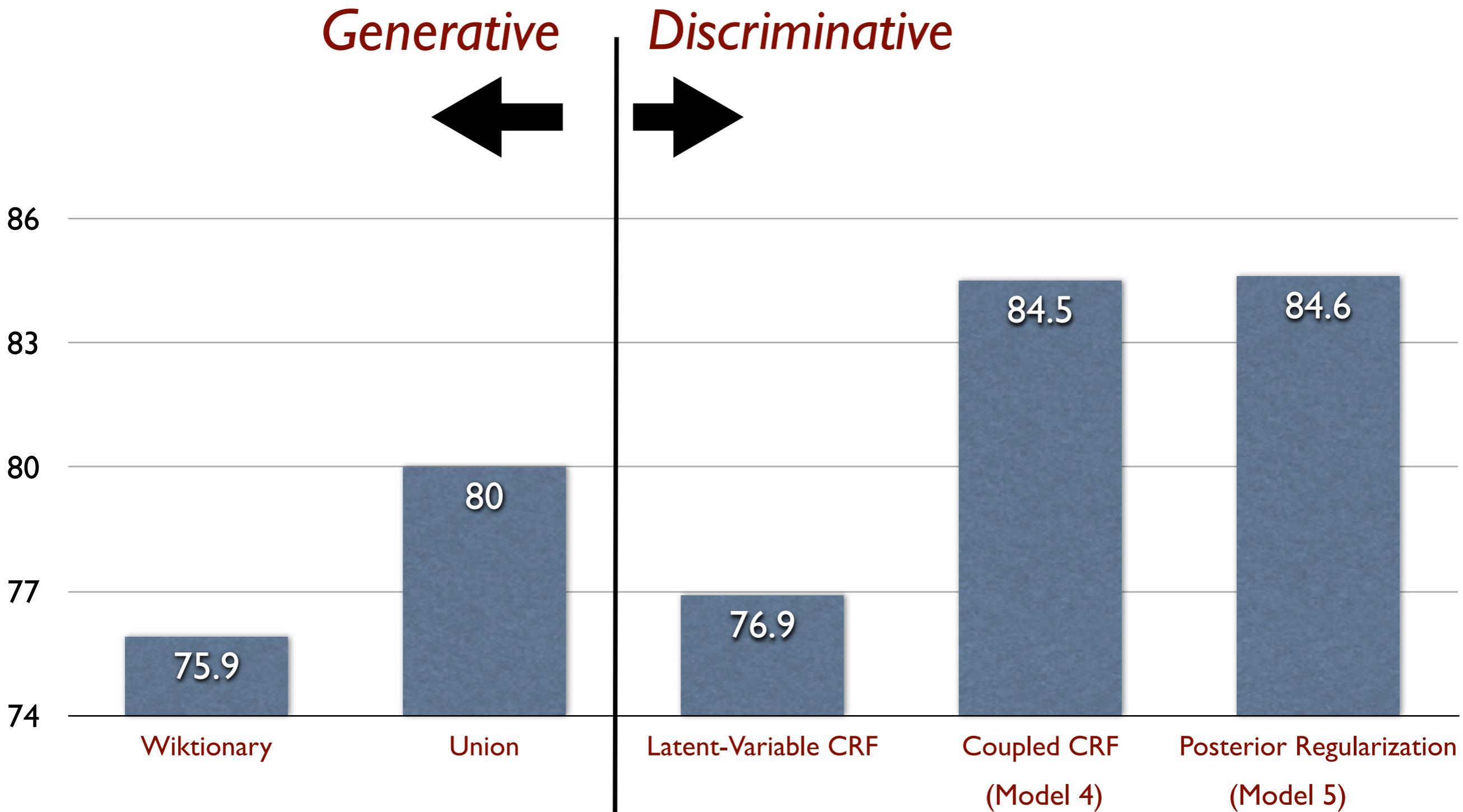
Measured across 15 Languages



Model Comparison



Measured across 15 Languages



Model Comparison

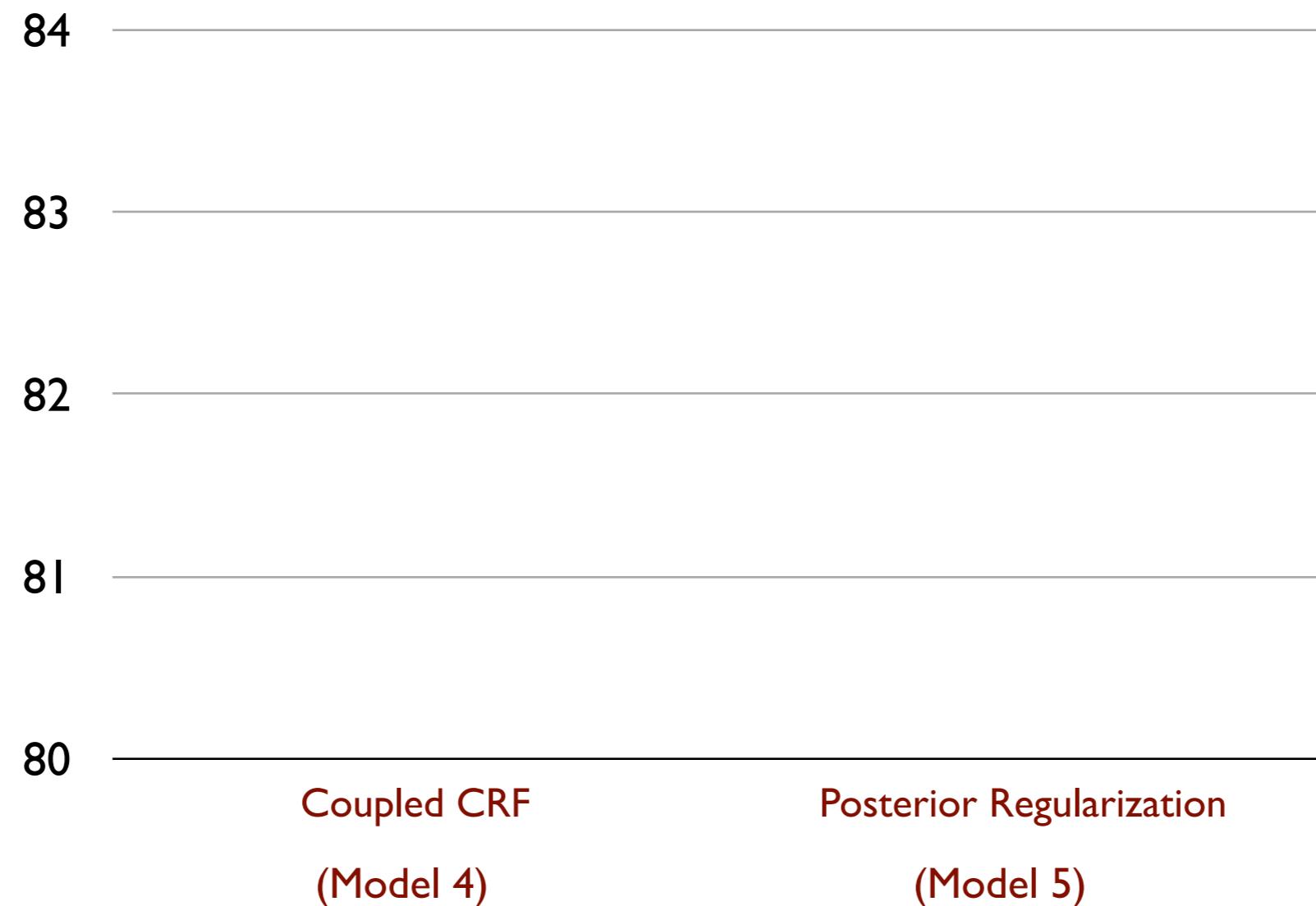


Measured across 17 Languages
(with Arabic and Hungarian)

Model Comparison



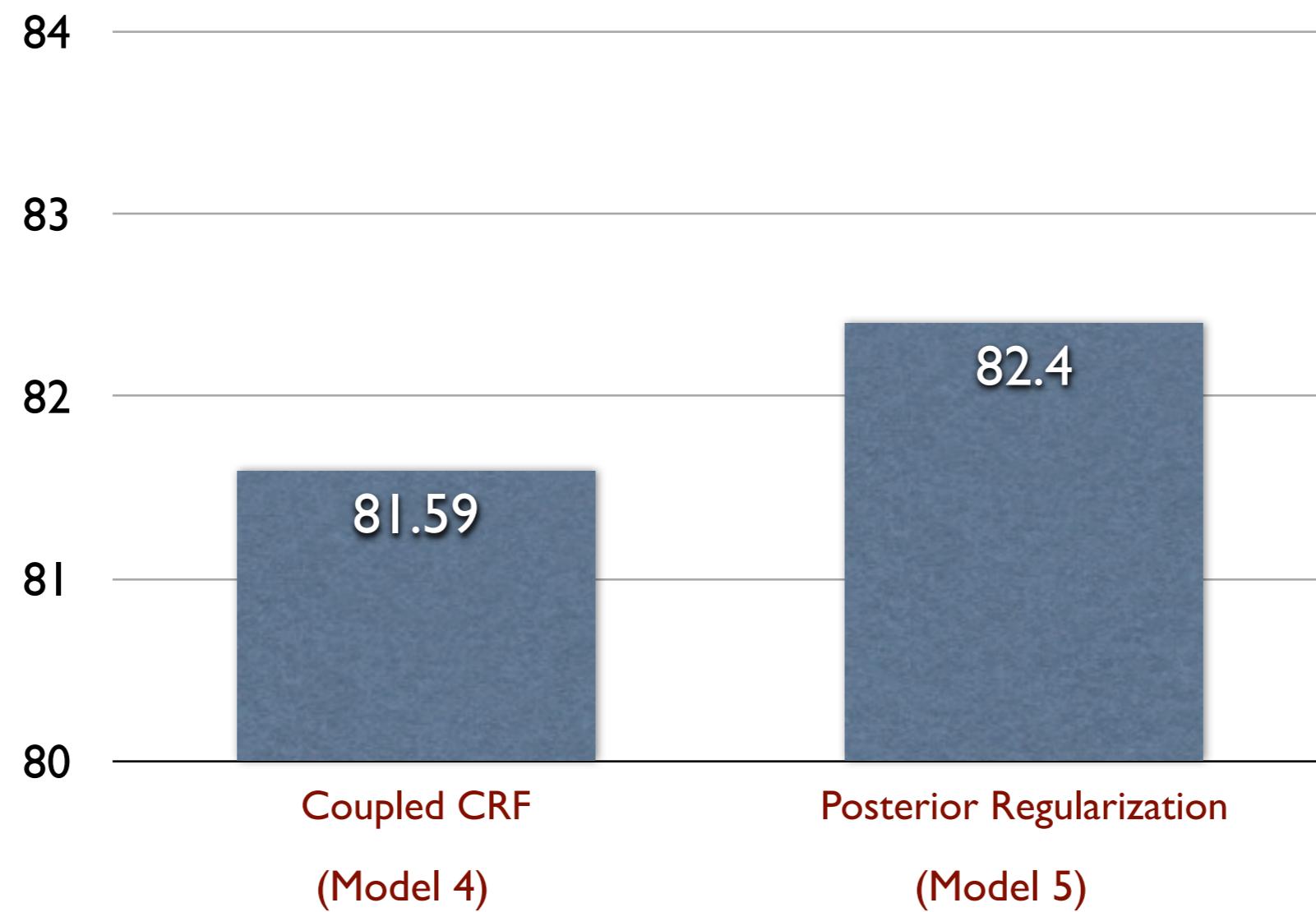
Measured across 17 Languages
(with Arabic and Hungarian)



Model Comparison



Measured across 17 Languages
(with Arabic and Hungarian)



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



dipanjand@google.com