Hierarchical Machine Translation

Wilker Aziz

Universiteit van Amsterdam
w.aziz@uva.nl

April 11, 2016

Content

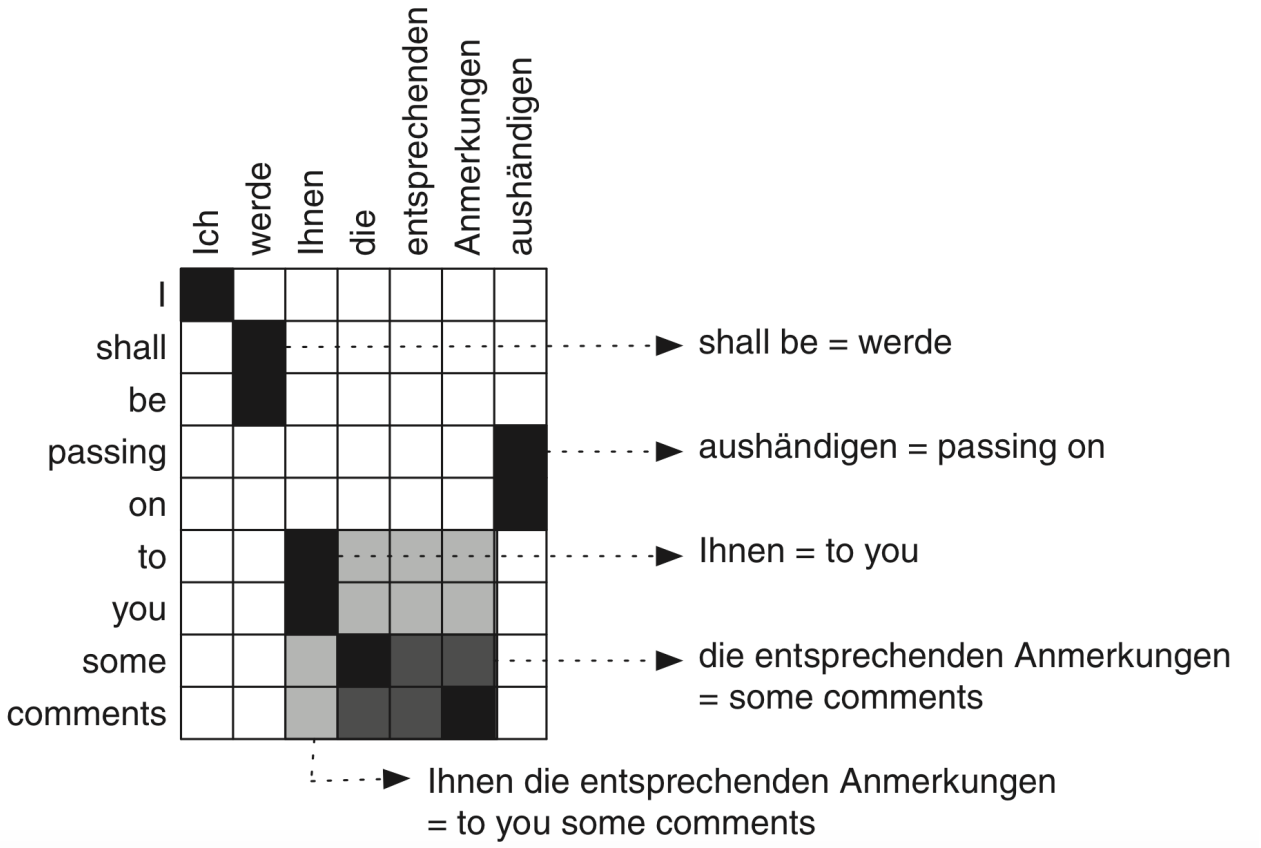


Figure : Koehn [2010]

werde X aushändigen | shall be passing on X

Why hierarchical structure?

Better generalisation

- compositionality
- reordering

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Reordering is arguably one of the hardest problems in MT

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Reordering is arguably one of the hardest problems in MT

- part of the model of translational equivalences
  the part that determines the space of translations

Key aspects

Expressiveness

- how much can two languages differ wrt word order?

Key aspects

Expressiveness

- how much can two languages differ wrt word order?

Modelling

- how many parameters do we have to estimate?

Content

Hierarchical phrase-based - Motivation

Local Reordering

	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

Hierarchical phrase-based - Motivation

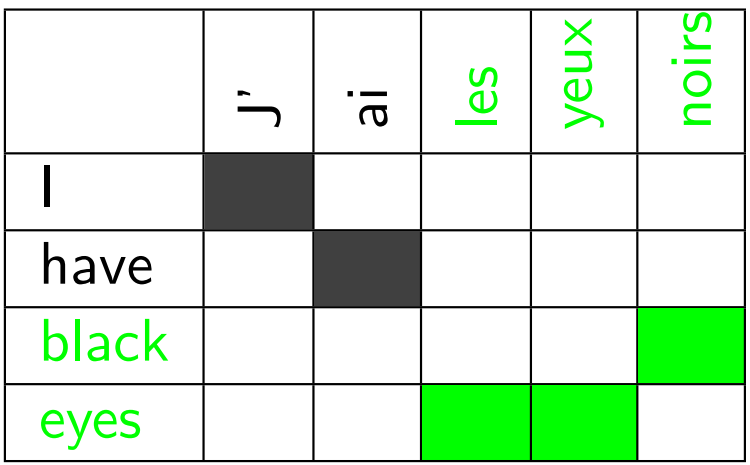
Local Reordering

	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- Monotone
  J'$_1$ ai$_2$ $\rightarrow$ I$_1$ have$_2$

Hierarchical phrase-based - Motivation

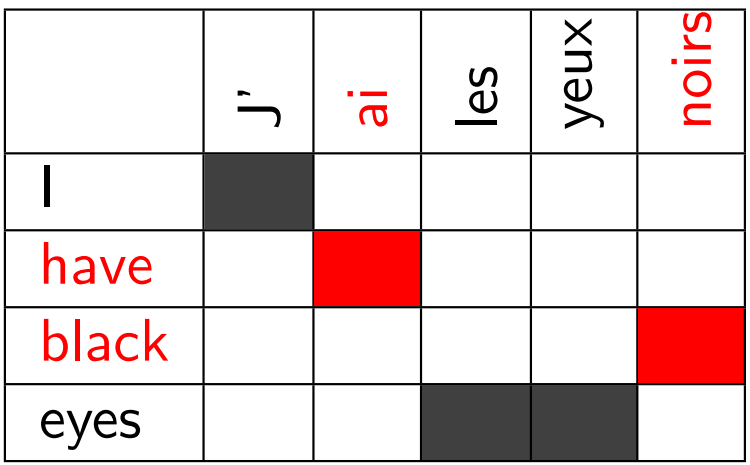
Local Reordering



- Swap
  les yeux$_4$ noirs$_5$ $\rightarrow$ black$_3$ eyes$_4$

Hierarchical phrase-based - Motivation

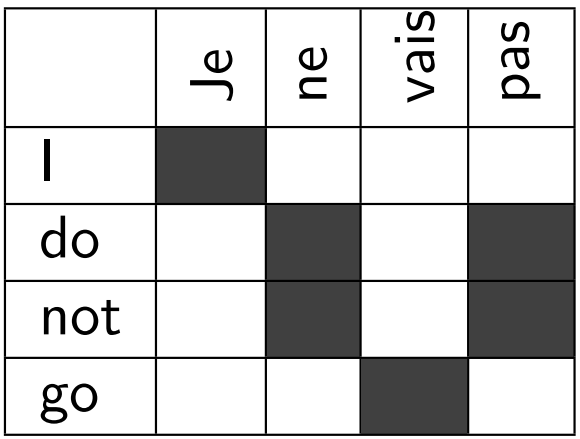
Local Reordering



- Discontinuous
  ai$_2$ X_{3-4} noirs$_5$ $\rightarrow$ have$_2$ black$_3$ X_4

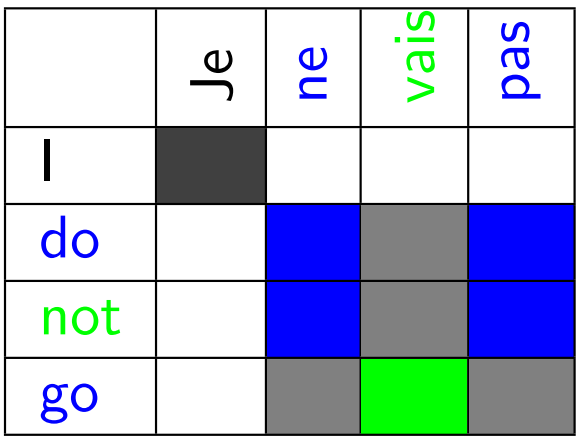
Hierarchical phrase-based - Motivation

Discontiguous Phrases



Hierarchical phrase-based - Motivation

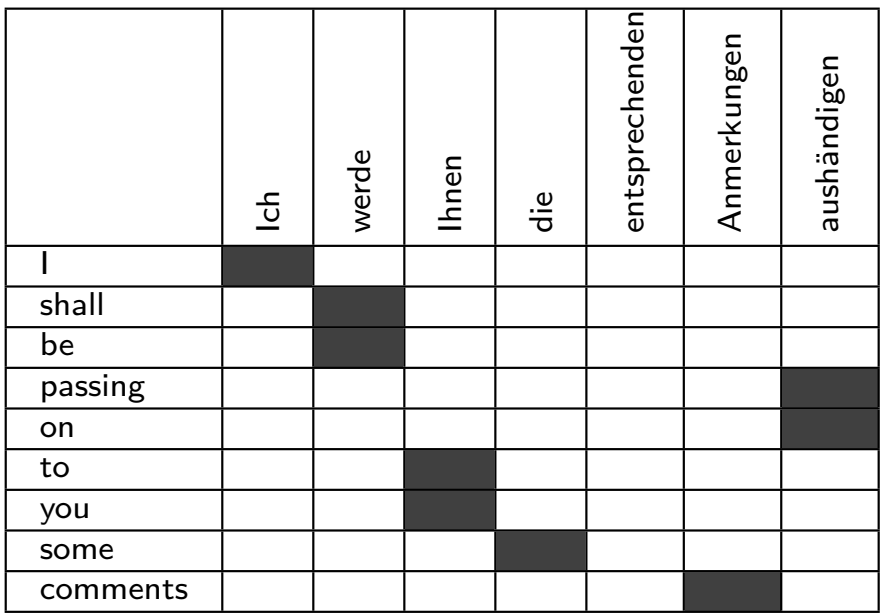
Discontiguous Phrases



- Gappy phrase
  ne vais pas → do not go
  ne X_{vais} pas → do not X_{go}

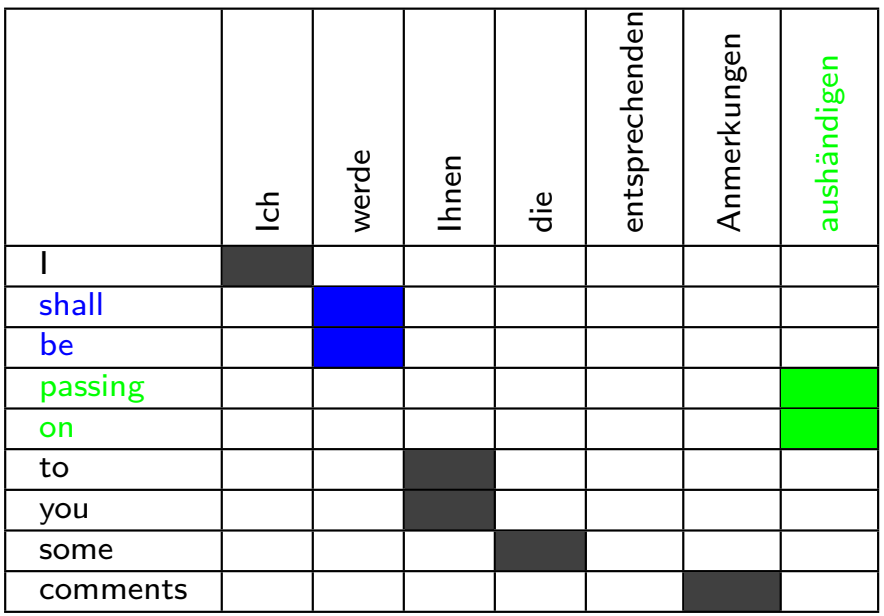
Hierarchical phrase-based - Motivation

Long Distance Reordering



Hierarchical phrase-based - Motivation

Long Distance Reordering



- How can we extract a biphrase for **shall be passing on**?

Hierarchical phrase-based - Motivation

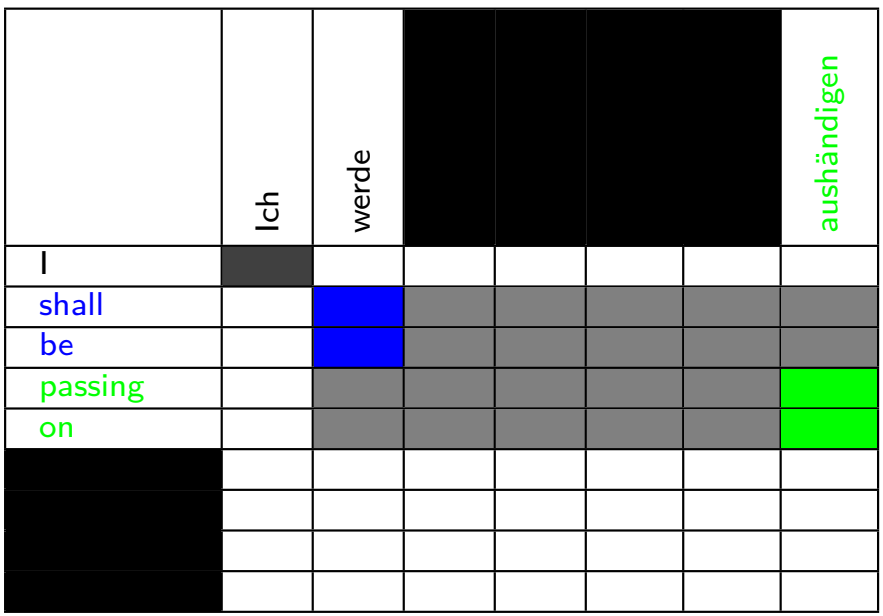
Long Distance Reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to			X				
you			X				
some				X			
comments						X	

- How can we extract a biphrase for **shall be passing on**?
- We cannot, we need to extract **to you some comments** along

Hierarchical phrase-based - Motivation

Long Distance Reordering



- How can we extract a biphrase for **shall be passing on**?
- We cannot, we need to extract **to you some comments** along
- Unless we replace all those words by a variable

Hierarchical phrase-based - Motivation

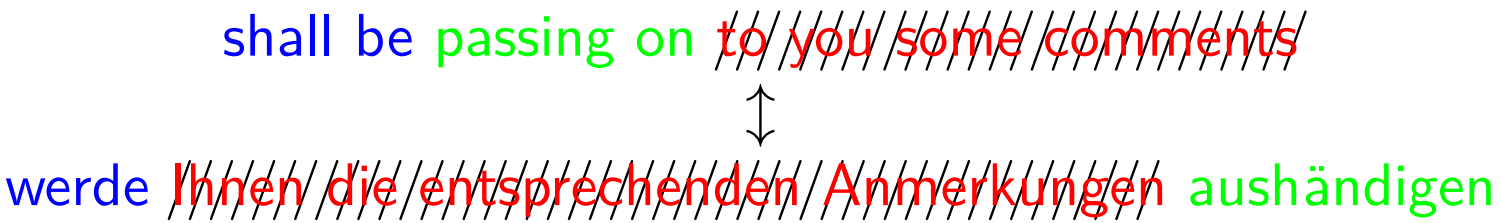
Long Distance Reordering

shall be passing on to you some comments

werde Ihnen die entsprechenden Anmerkungen aushändigen

Hierarchical phrase-based - Motivation

Long Distance Reordering



Hierarchical phrase-based - Motivation

Long Distance Reordering

shall be passing on X

$\updownarrow$

werde X aushändigen

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Motivation

- long-distance reordering

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Motivation

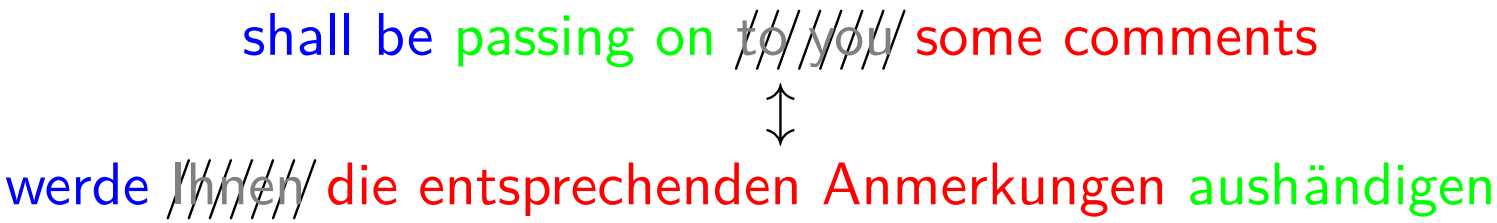
- long-distance reordering
- lexicalised reordering

Heuristic rule extraction

shall be passing on to you some comments

werde Ihnen die entsprechenden Anmerkungen aushändigen

Heuristic rule extraction

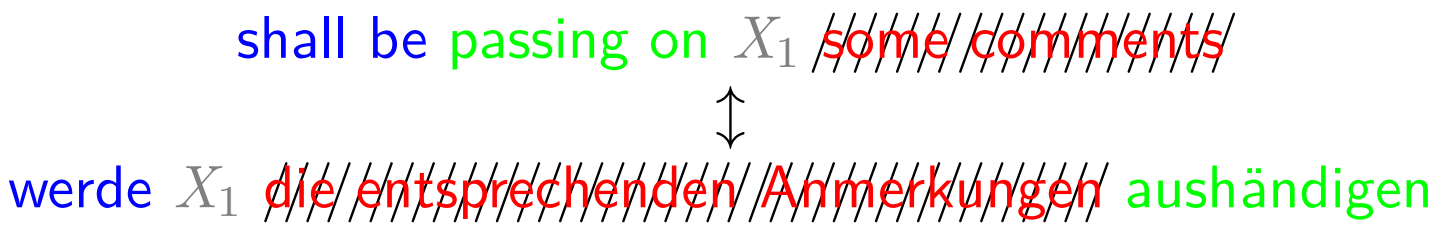


Heuristic rule extraction

shall be passing on X_1 some comments

werde X_1 die entsprechenden Anmerkungen aushändigen

Heuristic rule extraction



Heuristic rule extraction

shall be passing on X_1 X_2

$\updownarrow$

werde X_1 X_2 aushändigen

Heuristic rule extraction

$[X] \rightarrow$ shall be passing on X_1 X_2 | werde X_1 X_2 aushändigen

$[X] \rightarrow$ shall be passing on X_3 | werde X_3 aushändigen

$[X] \rightarrow$ to you | Ihnen

$[X] \rightarrow$ some comments | die entsprechenden Anmerkungen

$[X] \rightarrow$ to you some comments | Ihnen die entsprechenden Anmerkungen

Hiero - Constraints

Practical Limitations [Chiang, 2005]

Hiero - Constraints

Practical Limitations [Chiang, 2005]

- at most two nonterminal symbols

Hiero - Constraints

Practical Limitations [Chiang, 2005]

- at most two nonterminal symbols
- X spans at least 1 and at most 15 source words

Hiero - Constraints

Practical Limitations [Chiang, 2005]

- at most two nonterminal symbols
- X spans at least 1 and at most 15 source words
- no nonterminals next to each other in the source side

Hiero - Constraints

Practical Limitations [Chiang, 2005]

- at most two nonterminal symbols
- X spans at least 1 and at most 15 source words
- no nonterminals next to each other in the source side

les grandes maisons $\leftrightarrow$ the big houses
les X_1 maisons $\leftrightarrow$ the X_1 houses
les X_1 X_2 $\leftrightarrow$ the X_1 X_2
les X $\leftrightarrow$ the X

Hiero - Constraints

Practical Limitations [Chiang, 2005]

- at most two nonterminal symbols
- X spans at least 1 and at most 15 source words
- no nonterminals next to each other in the source side

$$\text{les grandes maisons} \leftrightarrow \text{the big houses}$$
$$\text{les } X_1 \text{ maisons} \leftrightarrow \text{the } X_1 \text{ houses}$$
$$\text{les } X_1 \ X_2 \leftrightarrow \text{the } X_1 \ X_2$$
$$\text{les } X \leftrightarrow \text{the } X$$

Glue rules

- $S \rightarrow \langle S_1 X_2, S_1 X_2 \rangle$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

$$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

$$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

  $$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

  $$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

  $$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

$$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

$$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

- Noisy-channel translation probability: $p(RHS_{source}|RHS_{target}, LHS)$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

  $$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

  $$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

  $$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

- Noisy-channel translation probability: $p(RHS_{source}|RHS_{target}, LHS)$

  $$p(\text{la maison } X_1|\text{the } X_1 \text{ house}, X)$$

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

  $$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

  $$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

  $$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

- Noisy-channel translation probability: $p(RHS_{source}|RHS_{target}, LHS)$

  $$p(\text{la maison } X_1|\text{the } X_1 \text{ house}, X)$$

- Lexical translation probability

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

$$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

$$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

- Noisy-channel translation probability: $p(RHS_{source}|RHS_{target}, LHS)$

$$p(\text{la maison } X_1|\text{the } X_1 \text{ house}, X)$$

- Lexical translation probability

$$\prod_{t_i \in RHS_{target}} p(t_i|RHS_{source}, a) \qquad \prod_{s_i \in RHS_{source}} p(s_i|RHS_{target}, a)$$

Hiero - Model

Log-linear combination of features

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \phi_i(\mathbf{d}, \mathbf{x})$$

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \phi_i(\mathbf{d}, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \prod_{r \in \mathbf{d}} \phi_i(r, \mathbf{x})$$

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \phi_i(\mathbf{d}, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \prod_{r \in \mathbf{d}} \phi_i(r, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \left(\sum_{r \in \mathbf{d}} \log \phi_i(r, \mathbf{x})\right)$$

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \phi_i(\mathbf{d}, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \prod_{r \in \mathbf{d}} \phi_i(r, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \left(\sum_{r \in \mathbf{d}} \log \phi_i(r, \mathbf{x})\right)$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_{r \in \mathbf{d}} \sum_i \lambda_i \log \phi_i(r, \mathbf{x})$$

Hiero - Model

Log-linear combination of features

$$p(\mathbf{d}, \mathbf{x}) = \prod_i \phi_i(\mathbf{d}, \mathbf{x})^{\lambda_i}$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \phi_i(\mathbf{d}, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \log \prod_{r \in \mathbf{d}} \phi_i(r, \mathbf{x})$$

$$\log p(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i \left(\sum_{r \in \mathbf{d}} \log \phi_i(r, \mathbf{x})\right)$$

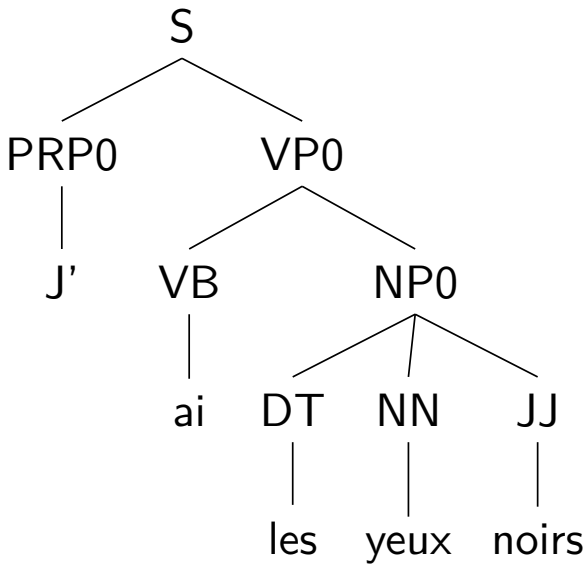
$$\log p(\mathbf{d}, \mathbf{x}) = \sum_{r \in \mathbf{d}} \sum_i \lambda_i \log \phi_i(r, \mathbf{x})$$

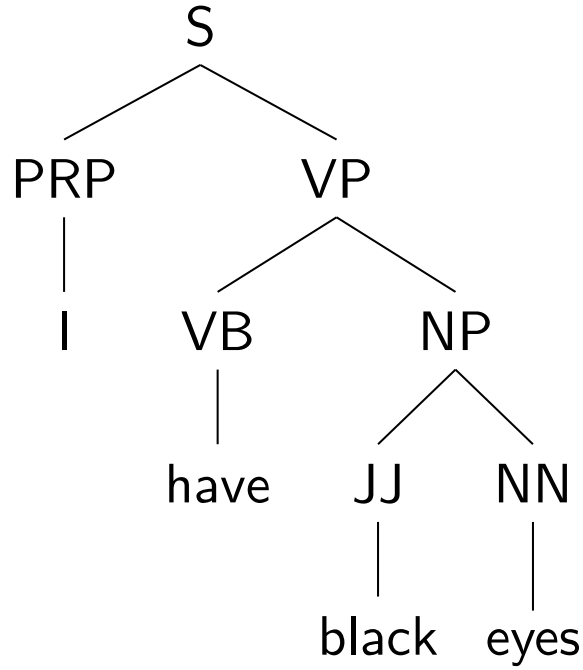
Linear model

$$f(\mathbf{d}, \mathbf{x}) = \sum_i \lambda_i h_i(\mathbf{d}, \mathbf{x}) = \sum_{r \in \mathbf{d}} \boldsymbol{\lambda}^\top \mathbf{h}(r, \mathbf{x})$$

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories



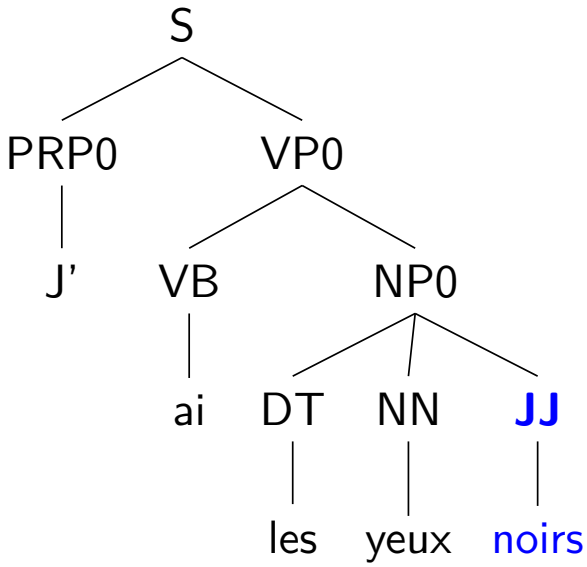


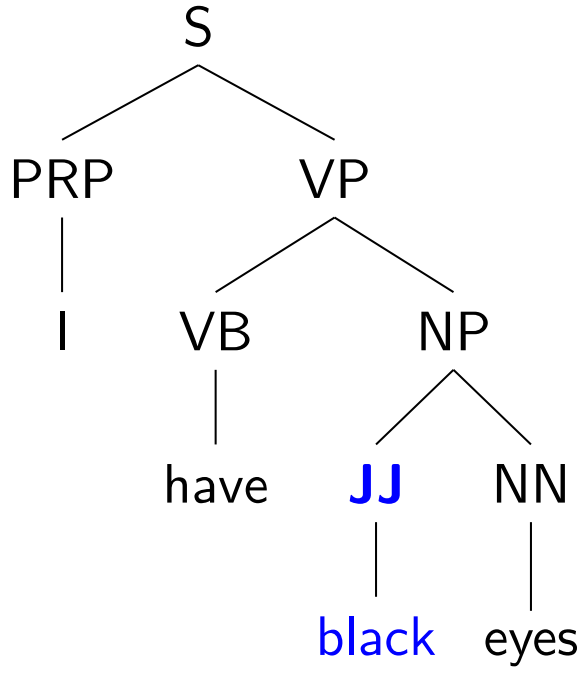
	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

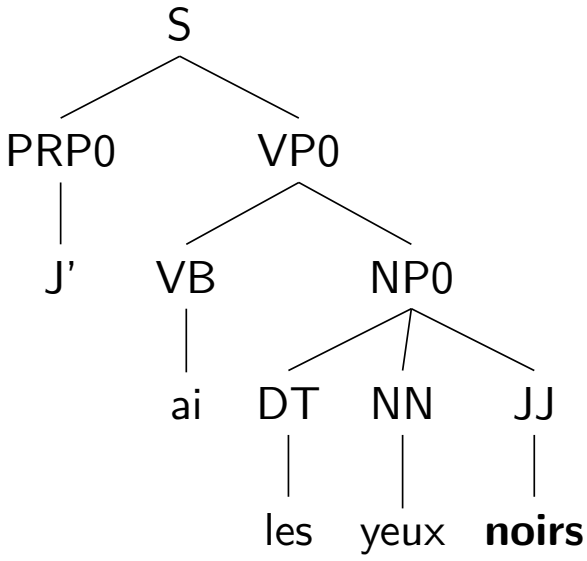
JJ → noirs | black

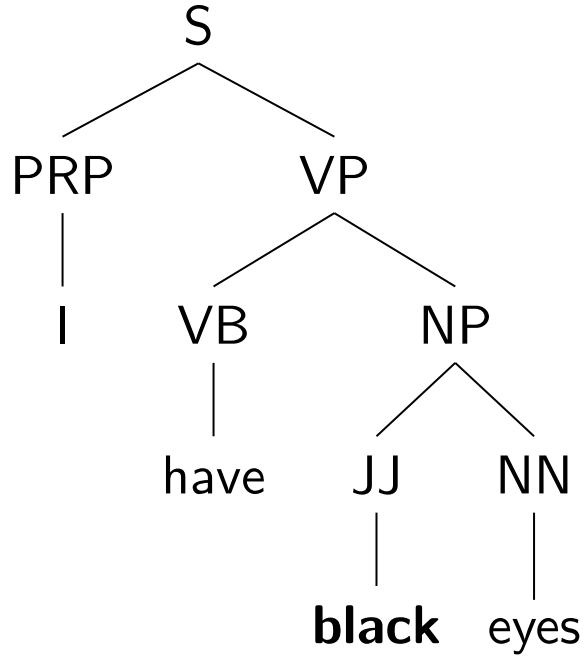
is straightforward

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories



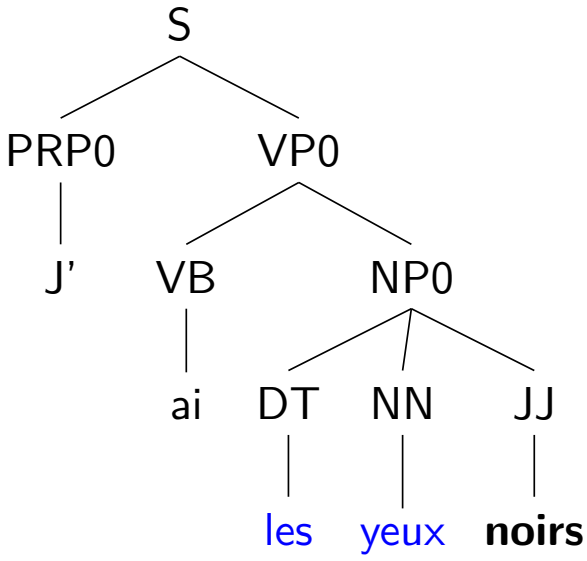


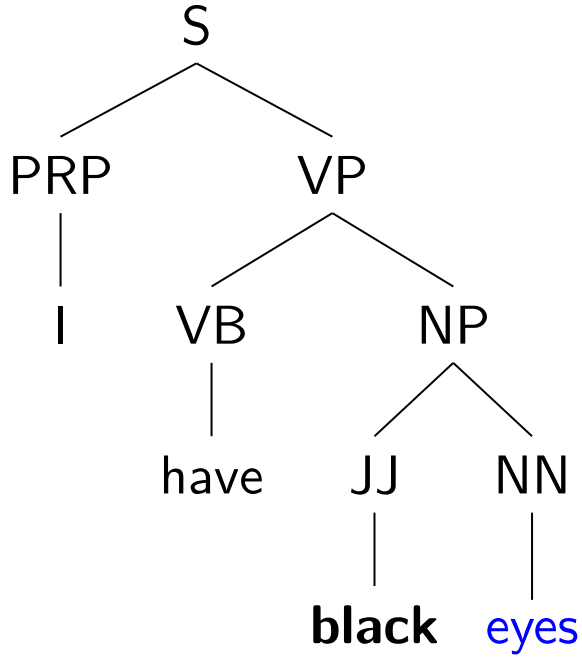
	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





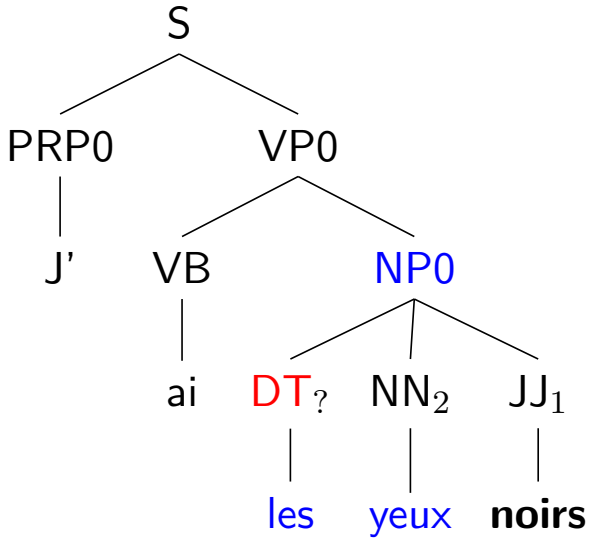
	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

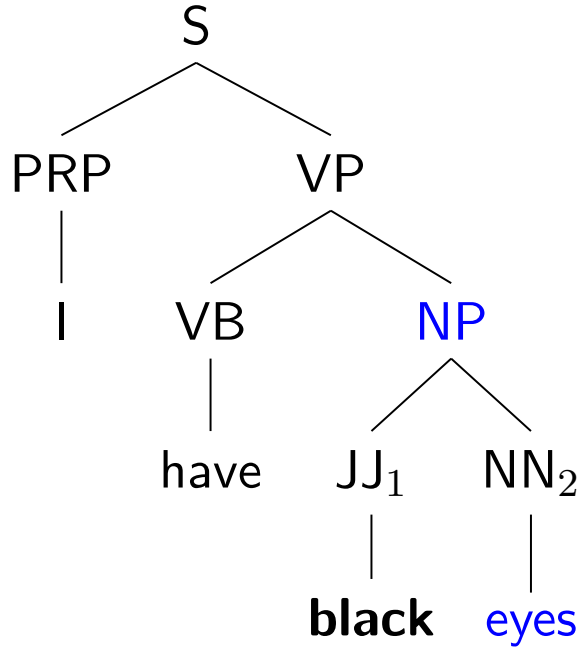
A single LHS $\rightarrow$ subtree

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

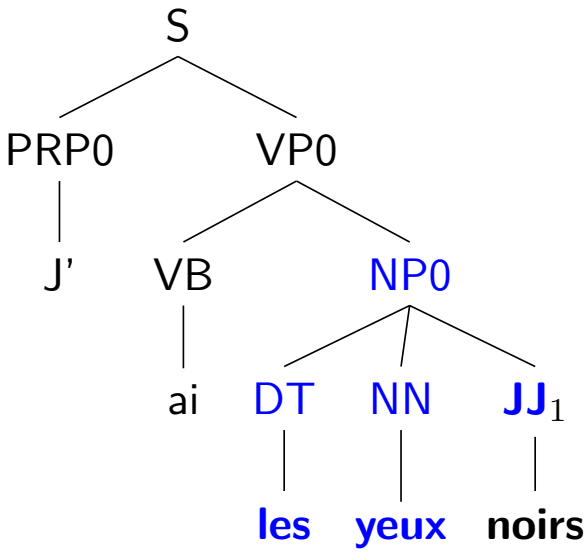
Use NP0/NP

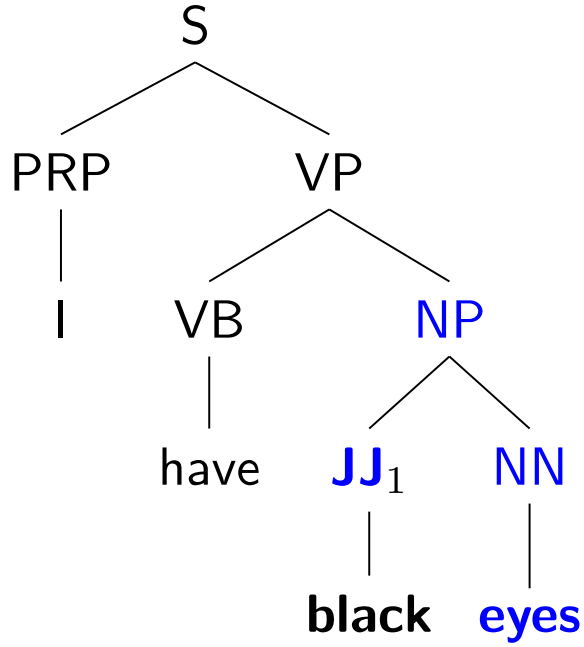
- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment

And constrained by syntactic categories





	J'	ai	les	yeux	noirs
I					
have					
black					JJ
eyes					

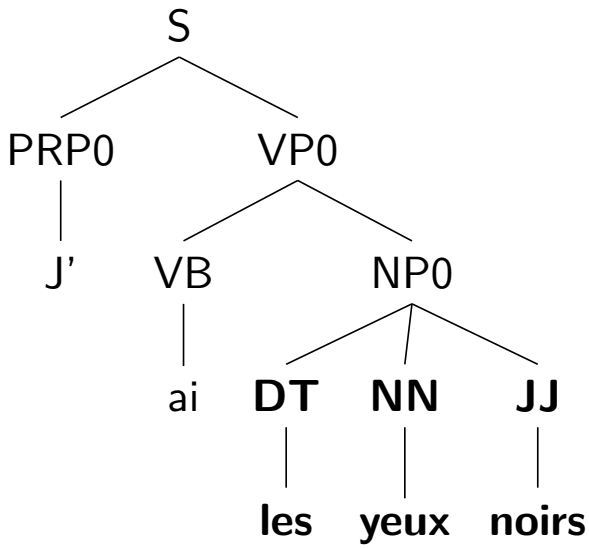
NP0/NP $\rightarrow$

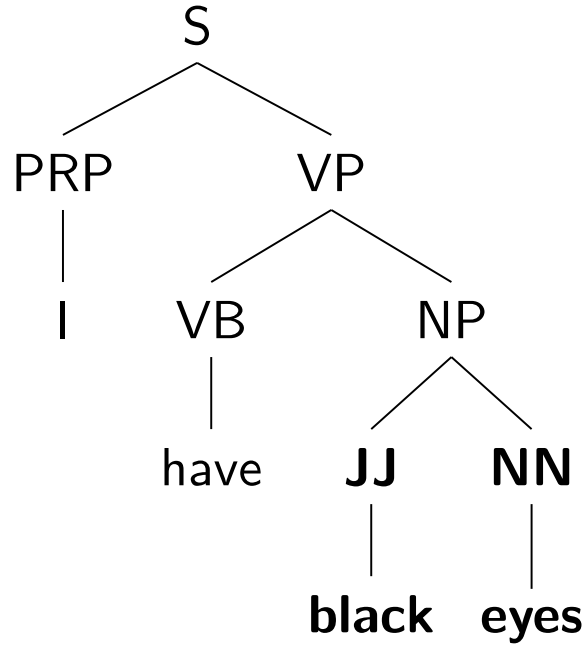
$\overset{DT}{les}\ \overset{NN}{yeux}$ JJ$_1$ | JJ$_1$ $\overset{NN}{eyes}$

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories



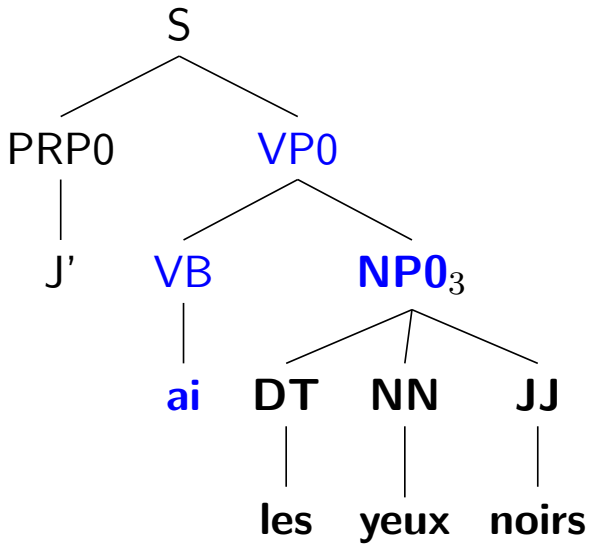


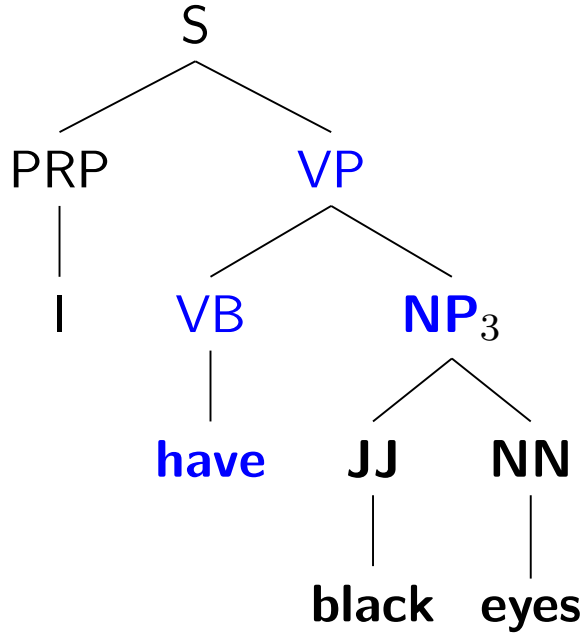
	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





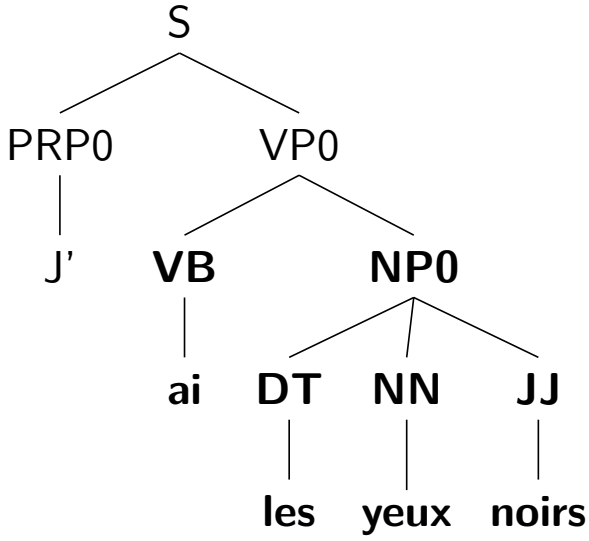
	J'	ai	les	yeux	noirs
I					
have					
black					NP
eyes			NP	NP	

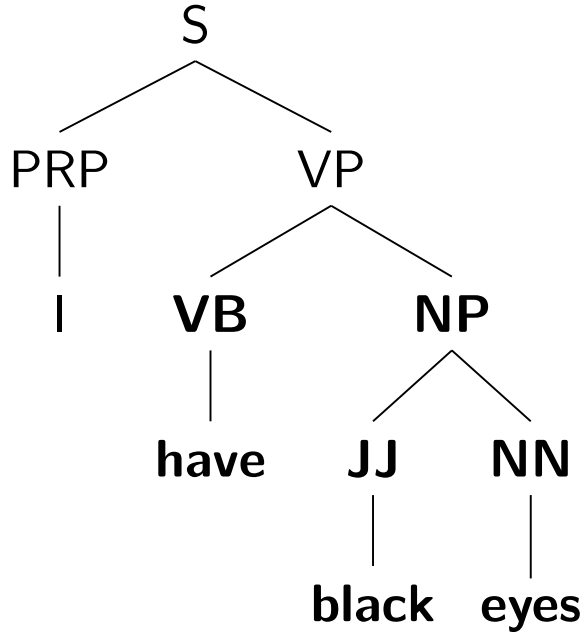
VP0/VP $\rightarrow$
$\overset{VB}{ai}$ NP0$_3$ | $\overset{VB}{have}$ NP$_3$

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories



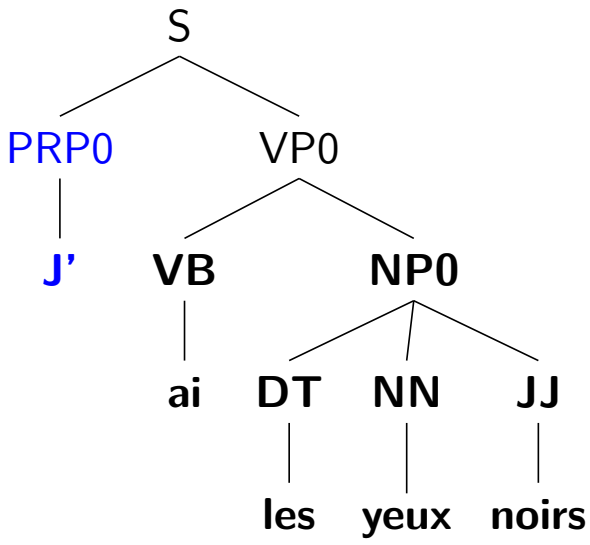


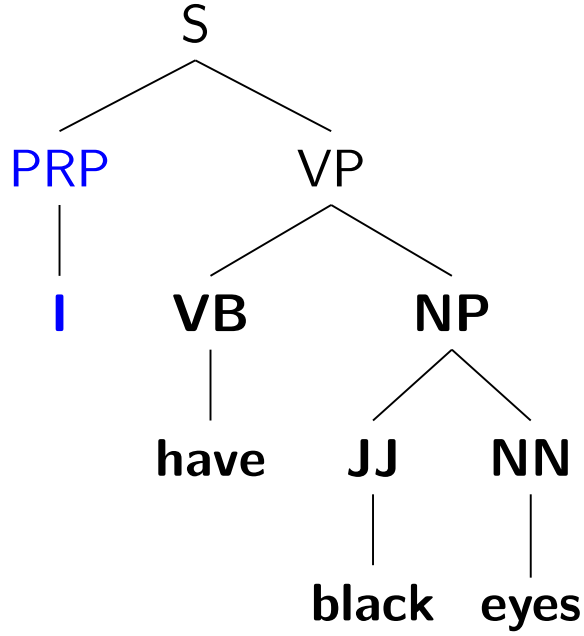
	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

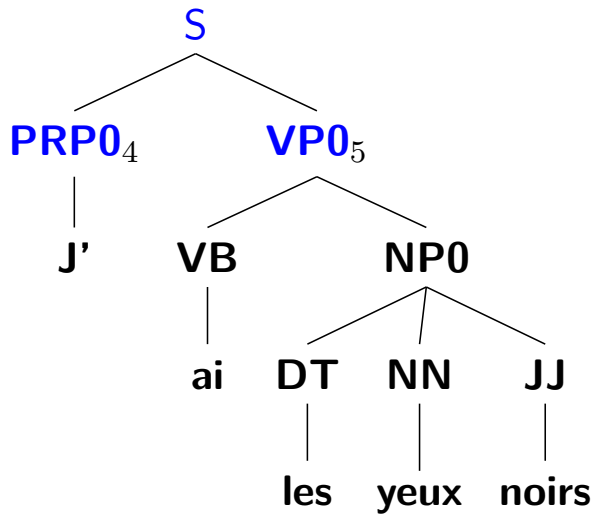
PRP0/PRP $\rightarrow$ J' | I

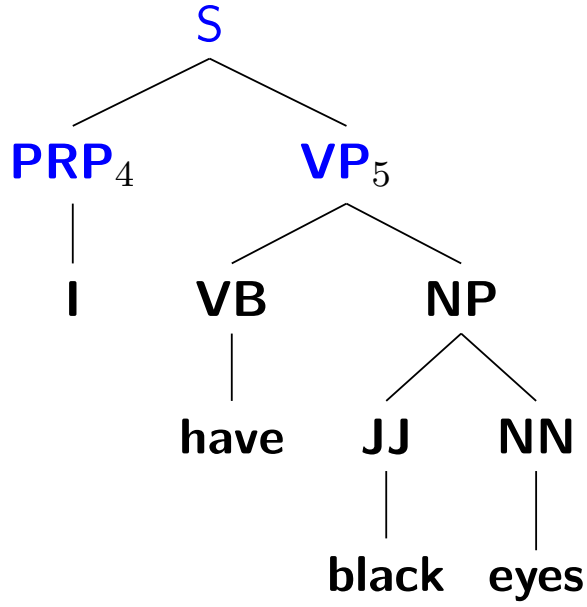
is straightforward

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Syntactic Constraints

Rules are learnt from the word-alignment
And constrained by syntactic categories





	J'	ai	les	yeux	noirs
I	PRP				
have		VP			
black					VP
eyes			VP	VP	

S $\rightarrow$ PRP0$_4$ VP$_5$ | PRP$_4$ VP$_5$

- A context-free rule requires a single LHS
- A rule must be consistent with word-alignment
- Nonterminals in the RHS must align one-to-one

Grammar

Grammar

PRP0/PRP $\rightarrow$ J' | I

JJ $\rightarrow$ noirs | black

NP0/NP $\rightarrow$ $\overset{DT}{les}$ $\overset{NN}{yeux}$ JJ | JJ

VP0/VP $\rightarrow$ $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP

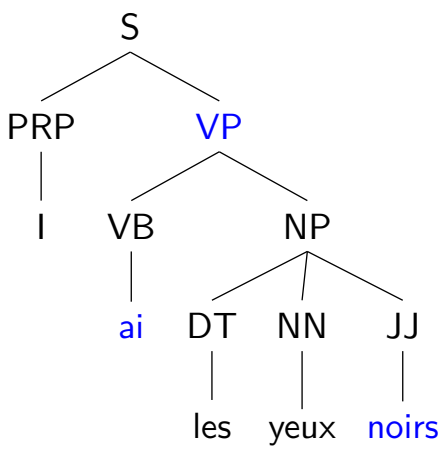
S $\rightarrow$ PRP0 VP0 | PRP VP

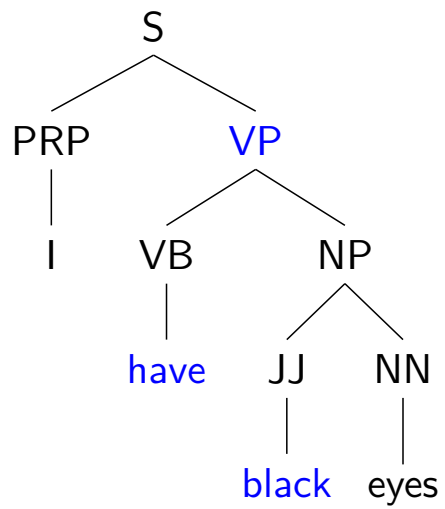
Syntax-based vs Hiero

More constraints on rules

Can we extract the discontiguous phrase **ai X noirs**?

Hiero: $X \rightarrow$ ai X_1 noirs | have black X_1
Syntactic: No!



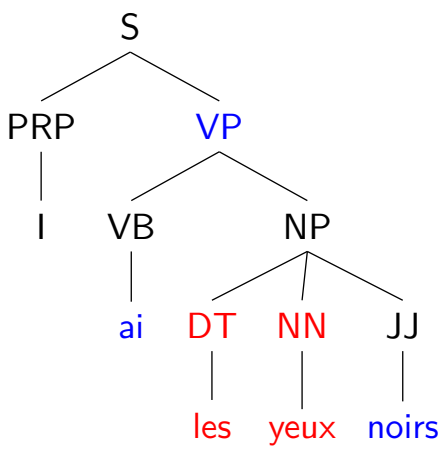


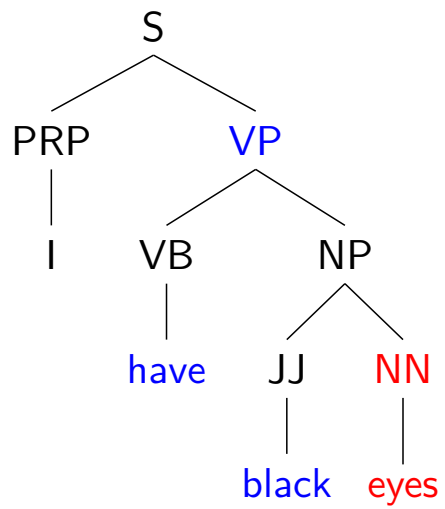
Syntax-based vs Hiero

More constraints on rules

Can we extract the discontiguous phrase **ai X noirs**?

Hiero: $X \rightarrow$ ai X_1 noirs | have black X_1
Syntactic: No!





Content

Decoding

Phrase-based

Tree-based

Decoding

Phrase-based

- Left-to-Right

Tree-based

- Bottom-Up

Decoding

Phrase-based

- Left-to-Right
- Beam Search

Tree-based

- Bottom-Up
- Chart Parsing

Decoding by Parsing

J' ai les yeux noirs

1. PRP0/PRP $\rightarrow$ J' | I
2. JJ $\rightarrow$ noirs | black
3. NP0/NP $\rightarrow$ $\overset{DT}{les}$ $\overset{NN}{yeux}$ JJ | JJ
4. VP0/VP $\rightarrow$ $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP
5. S $\rightarrow$ PRP0 VP0 | PRP VP

Decoding by Parsing

J'$_1$ ai les yeux noirs

PRP0$_1$

|

J'$_1$

PRP$_1$

|

I$_1$

1. PRP0/PRP $\rightarrow$ J' | I
2. JJ $\rightarrow$ noirs | black
3. NP0/NP $\rightarrow$ $\overset{DT}{les}$ $\overset{NN}{yeux}$ JJ | JJ
4. VP0/VP $\rightarrow$ $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP
5. S $\rightarrow$ PRP0 VP0 | PRP VP

$\{$I$_1\}$

Decoding by Parsing

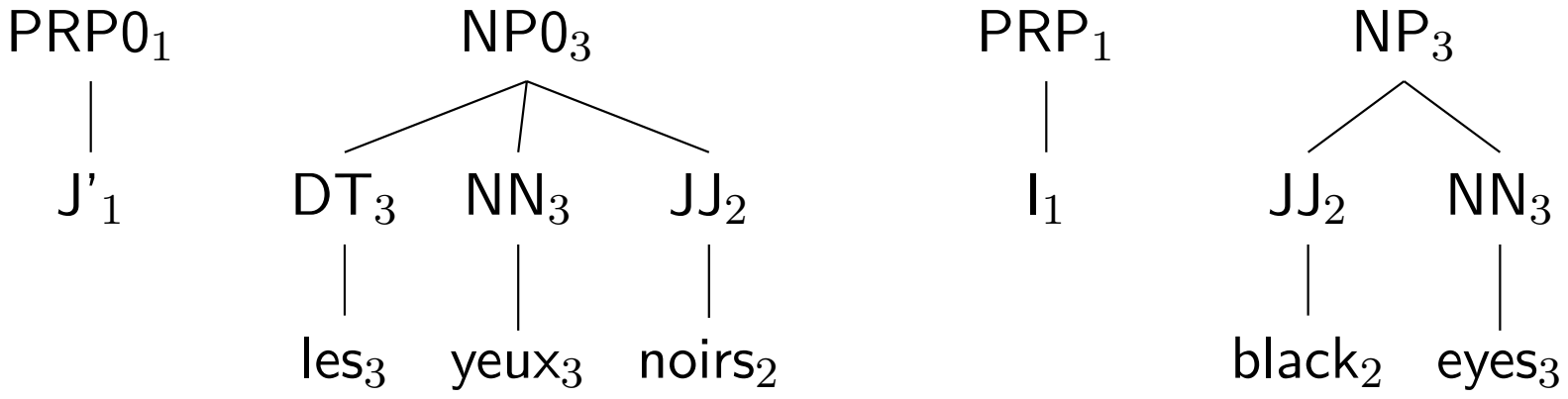
J'$_1$ ai les yeux noirs$_2$

1. PRP0/PRP → J' | I
2. JJ → noirs | black
3. NP0/NP → $\overset{DT}{les}\ \overset{NN}{yeux}$ JJ | JJ
4. VP0/VP → $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP
5. S → PRP0 VP0 | PRP VP

$\{$I$_1$,black$_2$$\}$

Decoding by Parsing

J'$_1$ ai les yeux$_3$ noirs$_2$

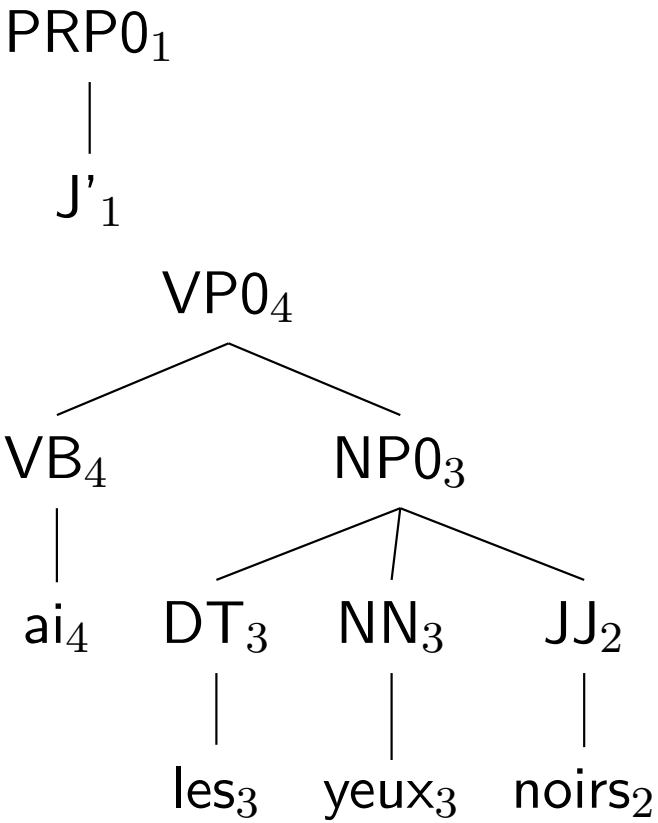


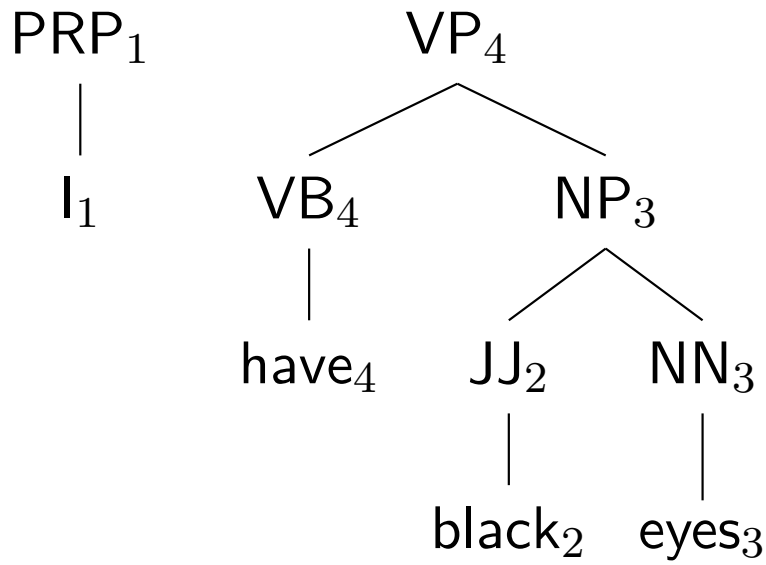
1. PRP0/PRP $\rightarrow$ J' | I
2. JJ $\rightarrow$ noirs | black
3. NP0/NP $\rightarrow$ $\overset{DT}{les}$ $\overset{NN}{yeux}$ JJ | JJ
4. VP0/VP $\rightarrow$ $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP
5. S $\rightarrow$ PRP0 VP0 | PRP VP

$\{$I$_1$, black$_2$ eyes$_3\}$

Decoding by Parsing

J'$_1$ ai$_4$ les yeux$_3$ noirs$_2$



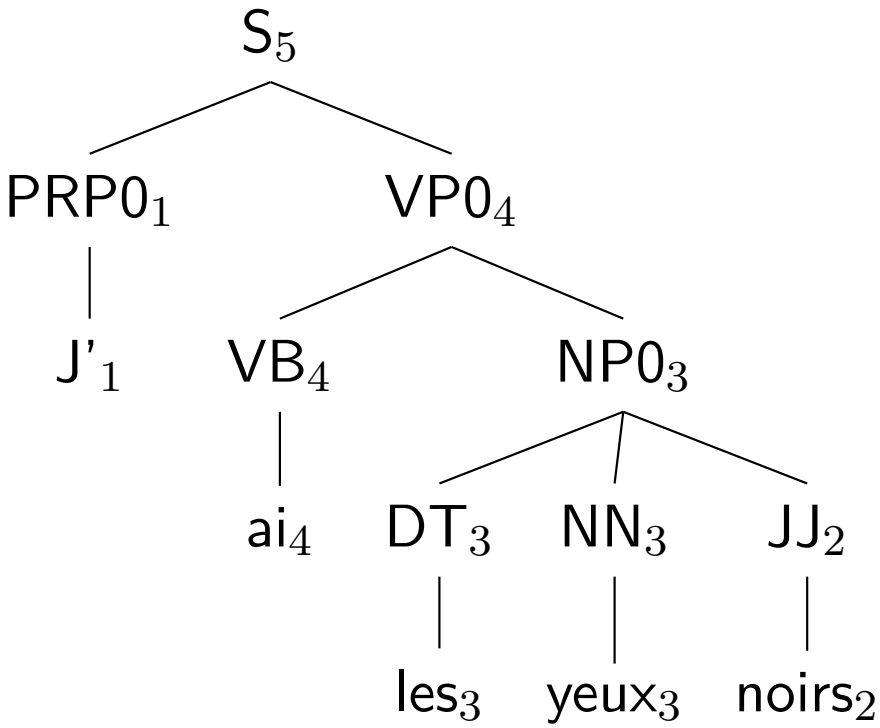


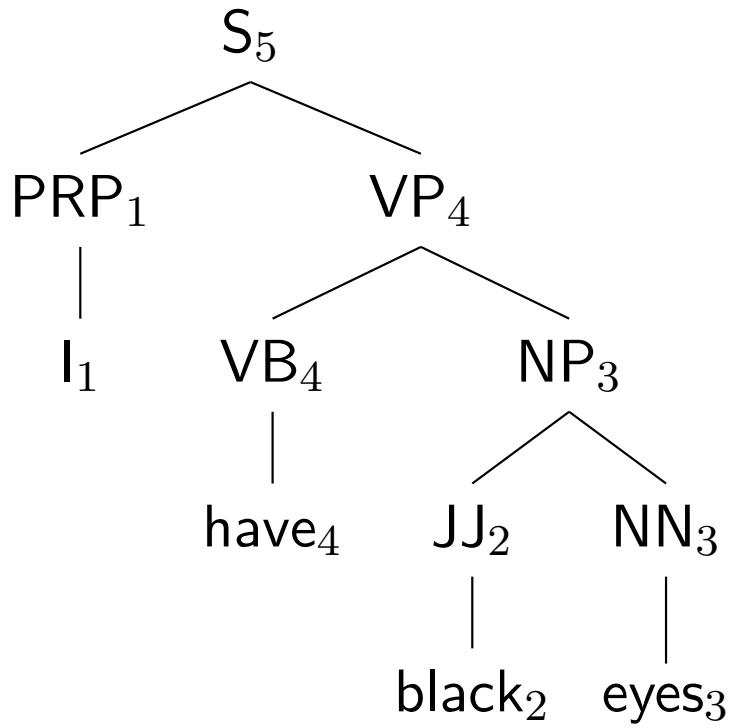
1. PRP0/PRP → J' | I
2. JJ → noirs | black
3. NP0/NP → $\overset{DT}{les}\ \overset{NN}{yeux}$ JJ | JJ
4. VP0/VP → $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP

{I$_1$, have$_4$ black$_2$ eyes$_3$}

Decoding by Parsing

J'$_1$ ai$_4$ les yeux$_3$ noirs$_2$





1. PRP0/PRP → J' | I
2. JJ → noirs | black
3. NP0/NP → $\overset{DT}{les}\ \overset{NN}{yeux}$ JJ | JJ
4. VP0/VP → $\overset{VB}{ai}$ NP0 | $\overset{VB}{have}$ NP
5. S → PRP0 VP0 | PRP VP

{I$_1$ have$_4$ black$_2$ eyes$_3$}

Conclusions and further reading

Hierarchical structure

- reasonable accounts of languages with different word-order

Conclusions and further reading

Hierarchical structure

- reasonable accounts of languages with different word-order
  - however, rather strict/fixed word-order (simpler morphology)

Conclusions and further reading

Hierarchical structure

- reasonable accounts of languages with different word-order
  - however, rather strict/fixed word-order (simpler morphology)

Linguistically-informed labels

- constrain hiero grammar [Zollmann and Venugopal, 2006]
- tree-based grammars [DeNeefe and Knight, 2009]
- feature-rich models [Chiang et al., 2009]
- tree transducers and EM training [Galley et al., 2006]

Questions?

Earley intersection

AXIOMS

$$\frac{}{[S' \to \bullet S, q, q]} \quad q \in I$$

GOAL

$$\left[S' \to S\bullet, q, r\right] \quad q \in I \wedge r \in F$$

SCAN

$$\frac{[X \to \alpha \bullet x\beta, q, s]}{[X \to \alpha x \bullet \beta]} \quad \langle s, x, r\rangle \in E$$

PREDICT

$$\frac{[X \to \alpha \bullet Y\beta, q, r]}{[Y \to \bullet\gamma, r, r]} \quad Y \to \gamma \in R$$

COMPLETE

$$\frac{[X \to \alpha \bullet Y\beta, q, s]\,[Y \to \gamma\bullet, s, r]}{[X \to \alpha\, Y_{s,r} \bullet \beta, q, r]} \quad X \neq S'$$

ACCEPT

$$\frac{[S' \to \bullet S, q, q]\,[S \to \gamma\bullet, q, r]}{[S' \to S_{q,r}\bullet, q, r]} \quad r \in F$$

References I

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219873. URL http://www.aclweb.org/anthology/P05-1033.

David Chiang, Kevin Knight, and Wei Wang. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N/N09/N09-1025.

References II

Steve DeNeefe and Kevin Knight. Synchronous tree adjoining machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 727–736, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-62-6. URL `http://dl.acm.org/citation.cfm?id=1699571.1699607`.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia, July 2006. Association for Computational

References III

Linguistics. doi: 10.3115/1220175.1220296. URL http://www.aclweb.org/anthology/P06-1121.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521874157, 9780521874151.

Andreas Zollmann and Ashish Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 138–141, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1654650.1654671.