Automatically Improved Category Labels for Syntax-Based Statistical Machine Translation

Greg Hanneman
Language Technologies Institute

Ph.D. Thesis Proposal
January 18, 2011

Thesis Committee
Alon Lavie (chair), Stephan Vogel,
Noah Smith, and David Chiang (USC)

Carnegie Mellon

Syntax-Based Statistical MT

- In this thesis, means MT systems that
  - Acquire syntax in a supervised manner from constituency parse trees
  - Model it with a synchronous context-free grammar (SCFG)

$$\text{NP} \rightarrow \text{DET}\ \ \text{N}\ \ \text{ADJ} \quad (\text{“la voiture bleue”})$$

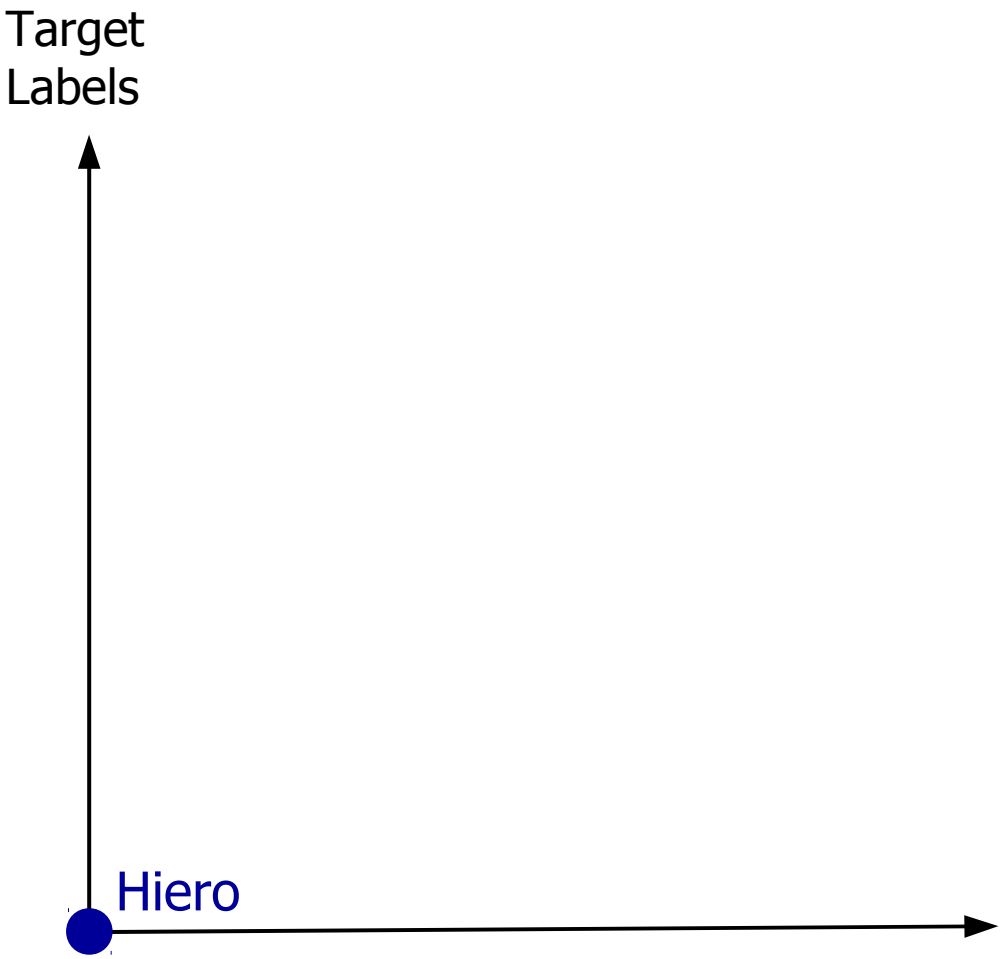
$$\text{NP} \rightarrow \text{DT}\ \ \text{JJ}\ \ \text{NN} \quad (\text{“the blue car”})$$

$$\text{NP::NP} \rightarrow [\text{DET}^1\ \text{N}^2\ \text{ADJ}^3]::[\text{DT}^1\ \text{JJ}^3\ \text{NN}^2]$$

SCFG Labeling Schemes

$$X::X \rightarrow [X^1 \text{ de } X^2]::[X^2\ X^1]$$

[Chiang 2005]



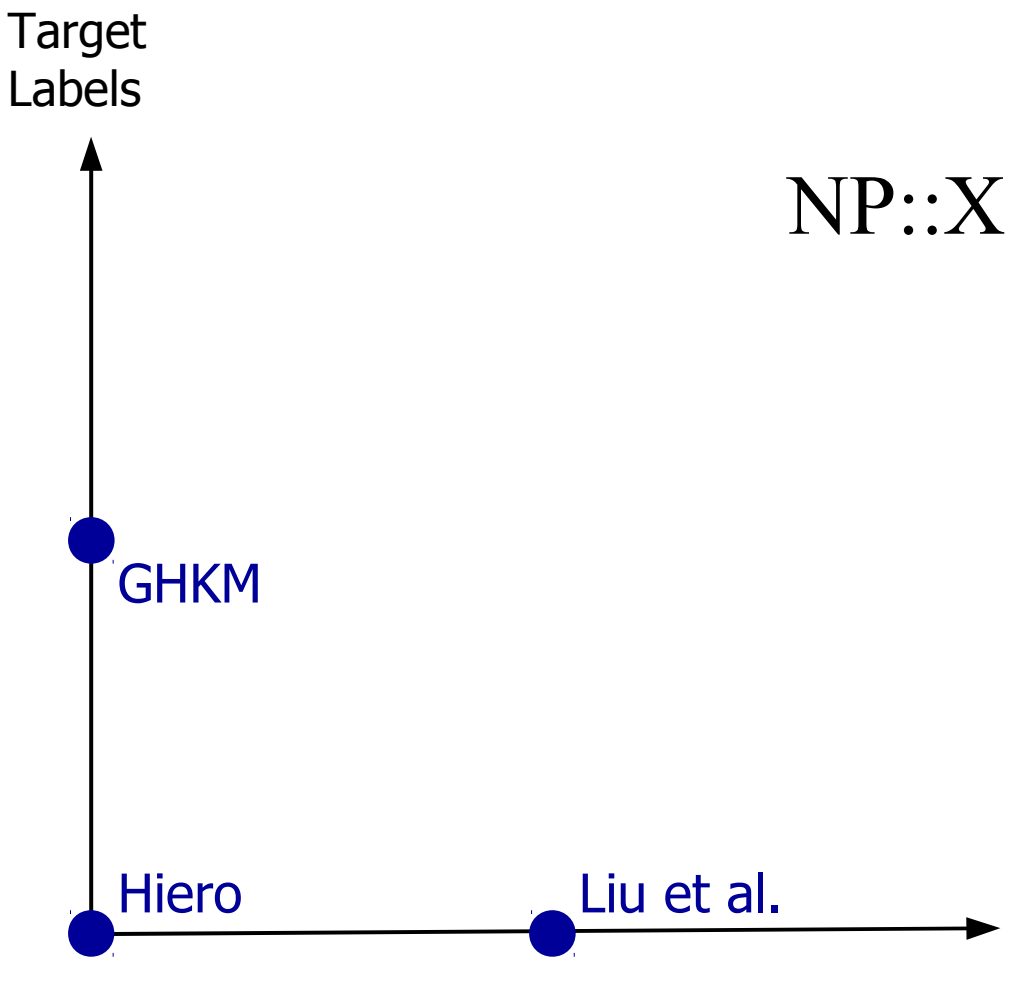
SCFG Labeling Schemes

$$\text{X::NP} \rightarrow [\text{X}^1 \text{ de } \text{X}^2]::[\text{N}^2 \text{ NP}^1]$$

[Galley et al. 2004]

$$\text{NP::X} \rightarrow [\text{NP}^1 \text{ de } \text{N}^2]::[\text{X}^2 \text{ X}^1]$$

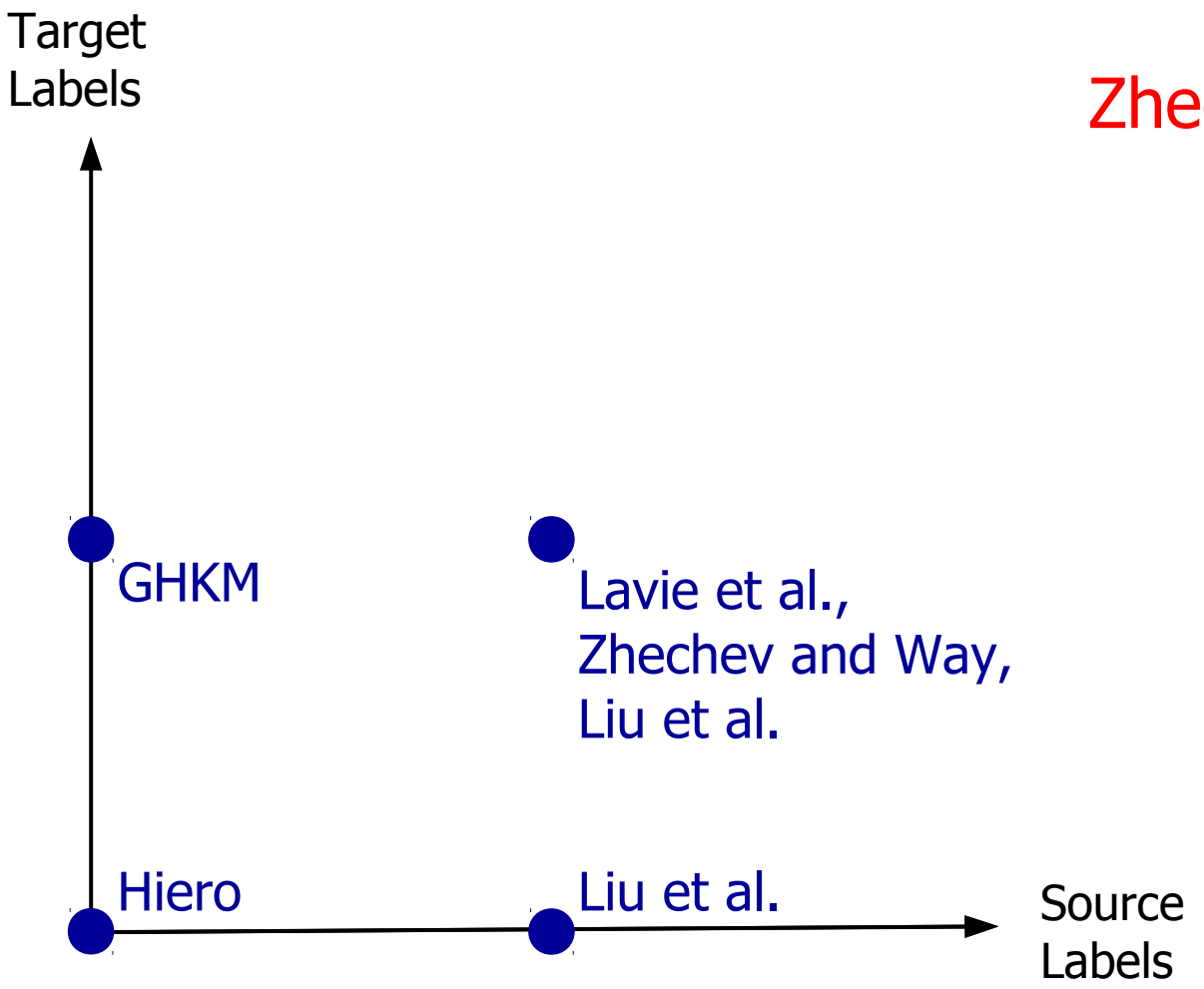
[Liu et al. 2006]



SCFG Labeling Schemes

$$\text{NP::NP} \rightarrow [\text{NP}^1 \text{ de } \text{N}^2]::[\text{N}^2 \text{ NP}^1]$$

[Lavie et al. 2008;
Zhechev and Way 2008;
Liu et al. 2009]

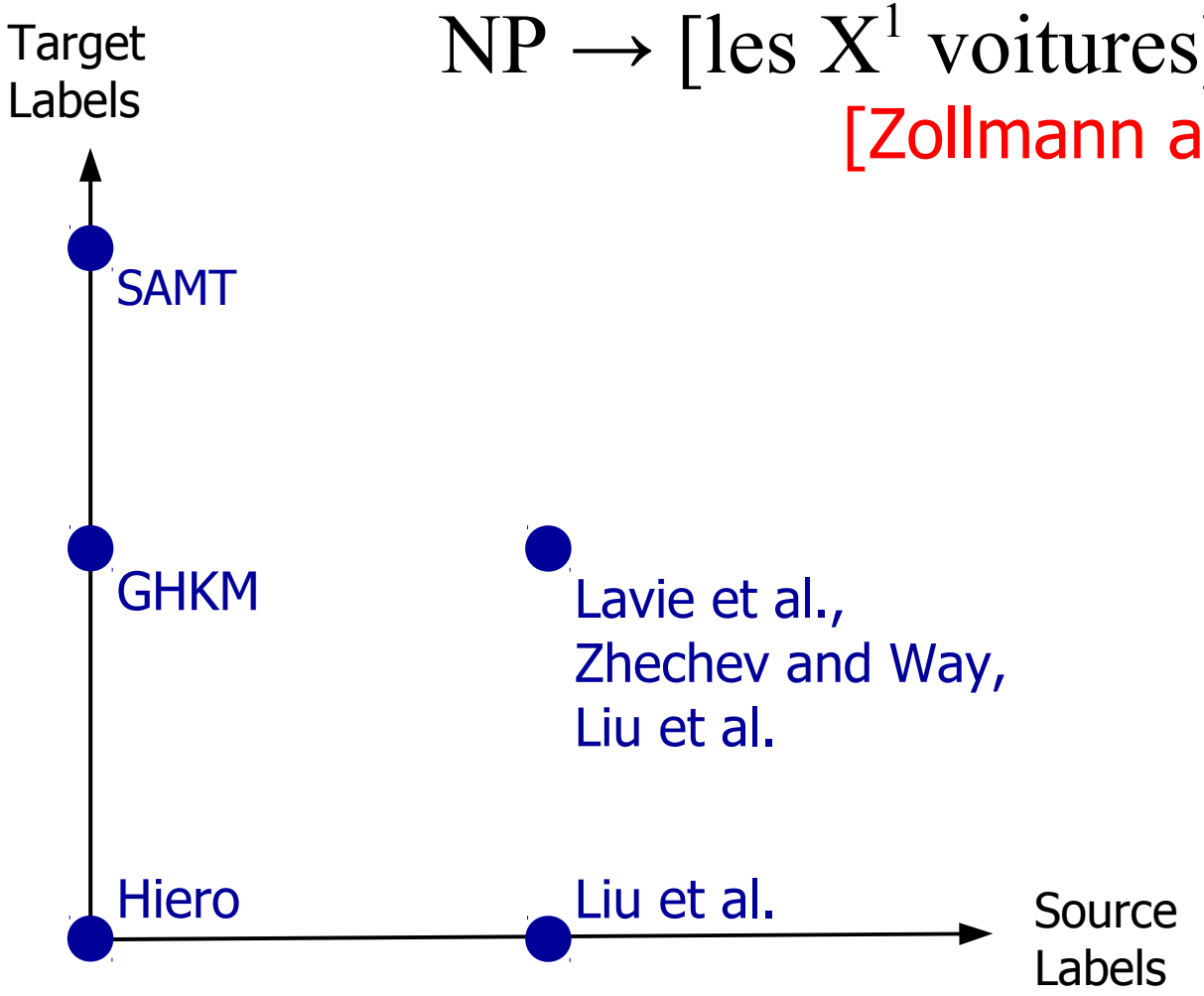


SCFG Labeling Schemes

$$S \rightarrow [\text{il y a } X^{1}]::[\text{there is } S/NP^{1}]$$

$$NP \rightarrow [\text{les } X^{1} \text{ voitures}]::[\text{the } JJ{+}JJ^{1} \text{ cars}]$$

[Zollmann and Venugopal 2006]

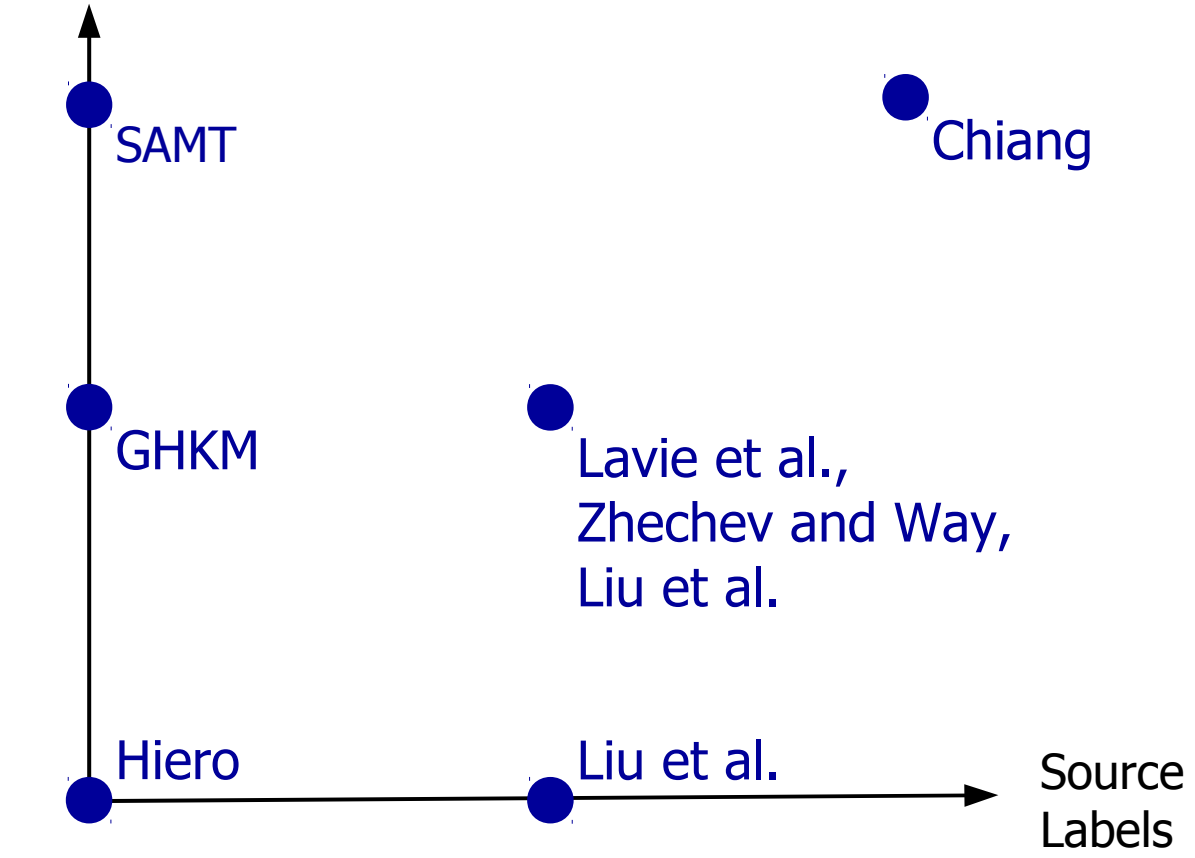


SCFG Labeling Schemes

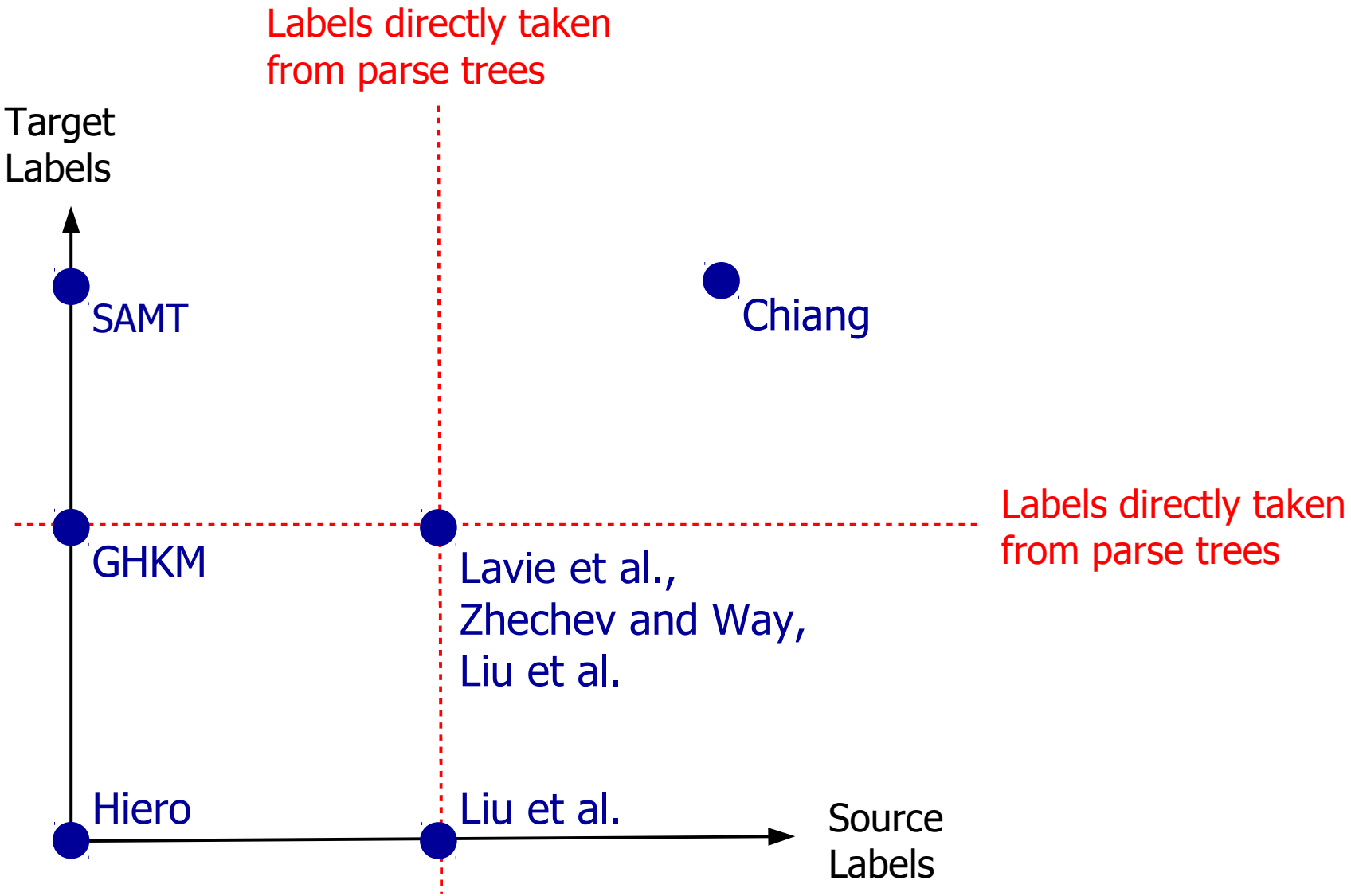
$$\text{S::S} \rightarrow [\text{il y a S/NP}^1]::[\text{there is S/NP}^1]$$

$$\text{NP::NP} \rightarrow [\text{les D+A}^1 \text{ voitures}]::[\text{the JJ+JJ}^1 \text{ cars}]$$

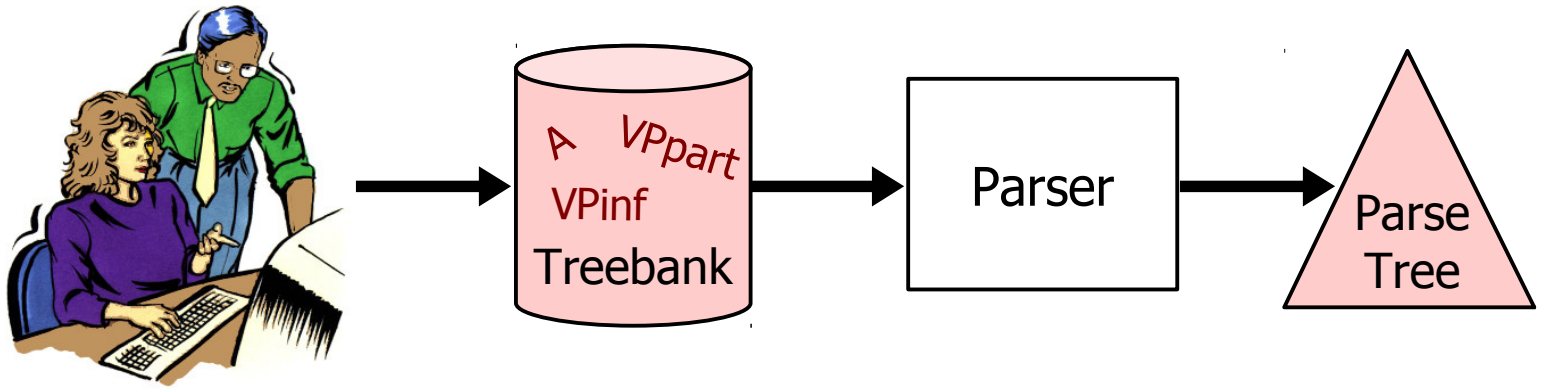
[Chiang 2010]



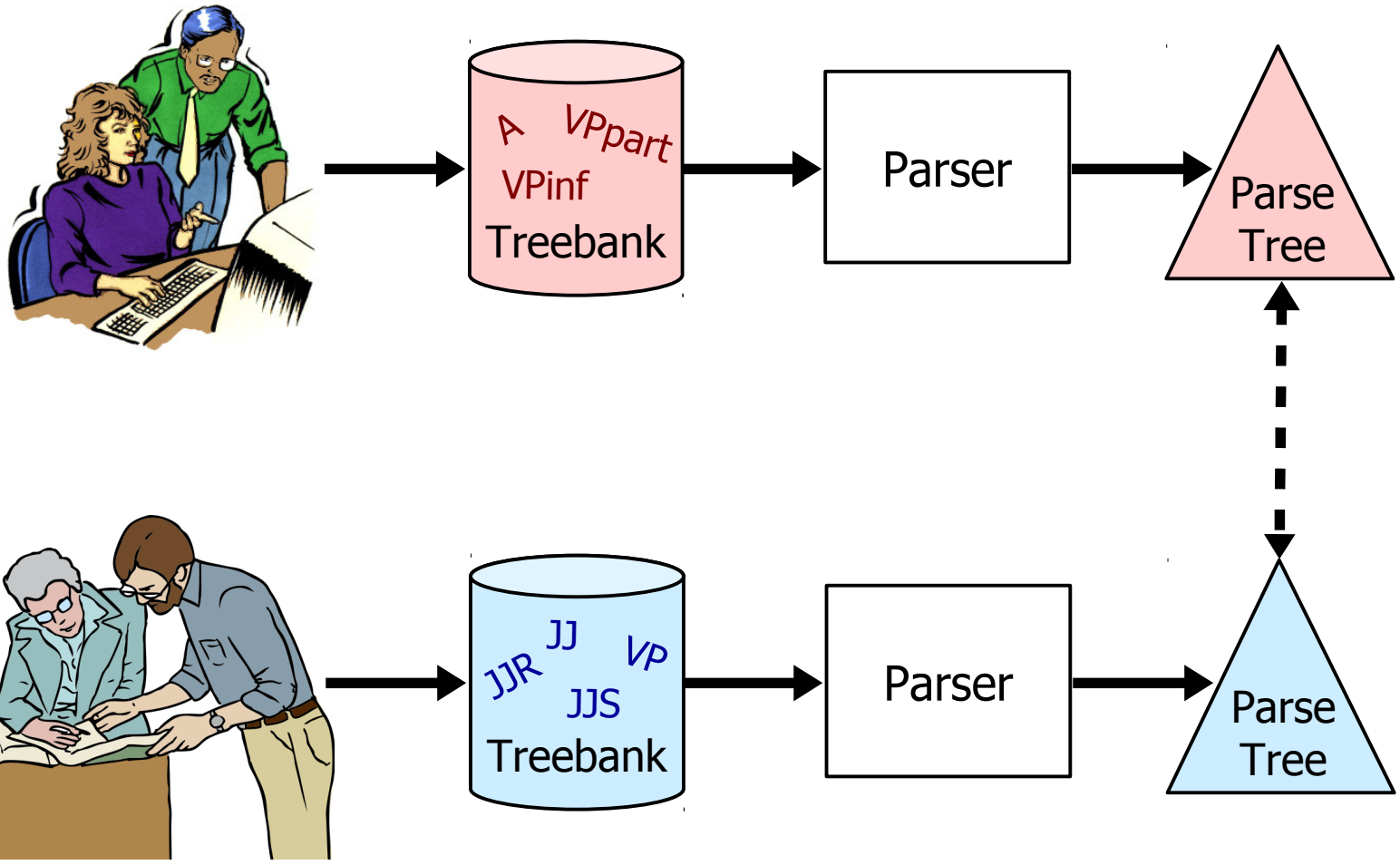
SCFG Labeling Schemes



Parse Tree Labels Assume a Lot



Parse Tree Labels Assume a Lot

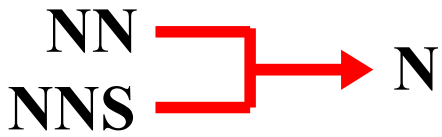


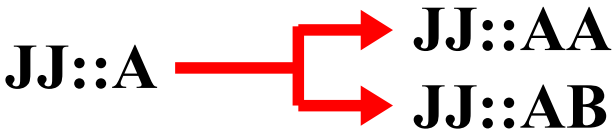
Thesis Statement

The nonterminal set used in SCFG-based MT is important.

Using category labels from treebanks is suboptimal for MT, but we can make labels better automatically.

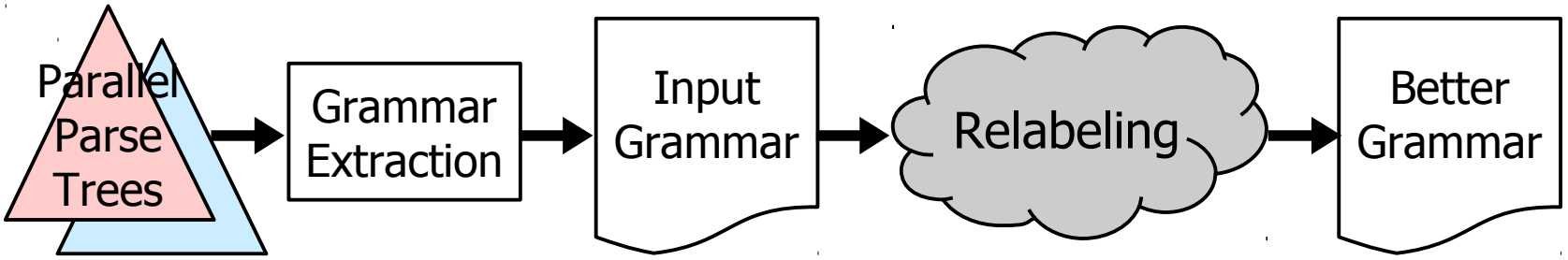
Thesis Proposal

- Define and measure the effect labels have
  - Spurious ambiguity, rule sparsity, and reordering precision
- Explore the space of labeling schemes
  - Collapsing labels 

  - Refining labels 

  - Correcting local labeling errors 

Thesis Claims

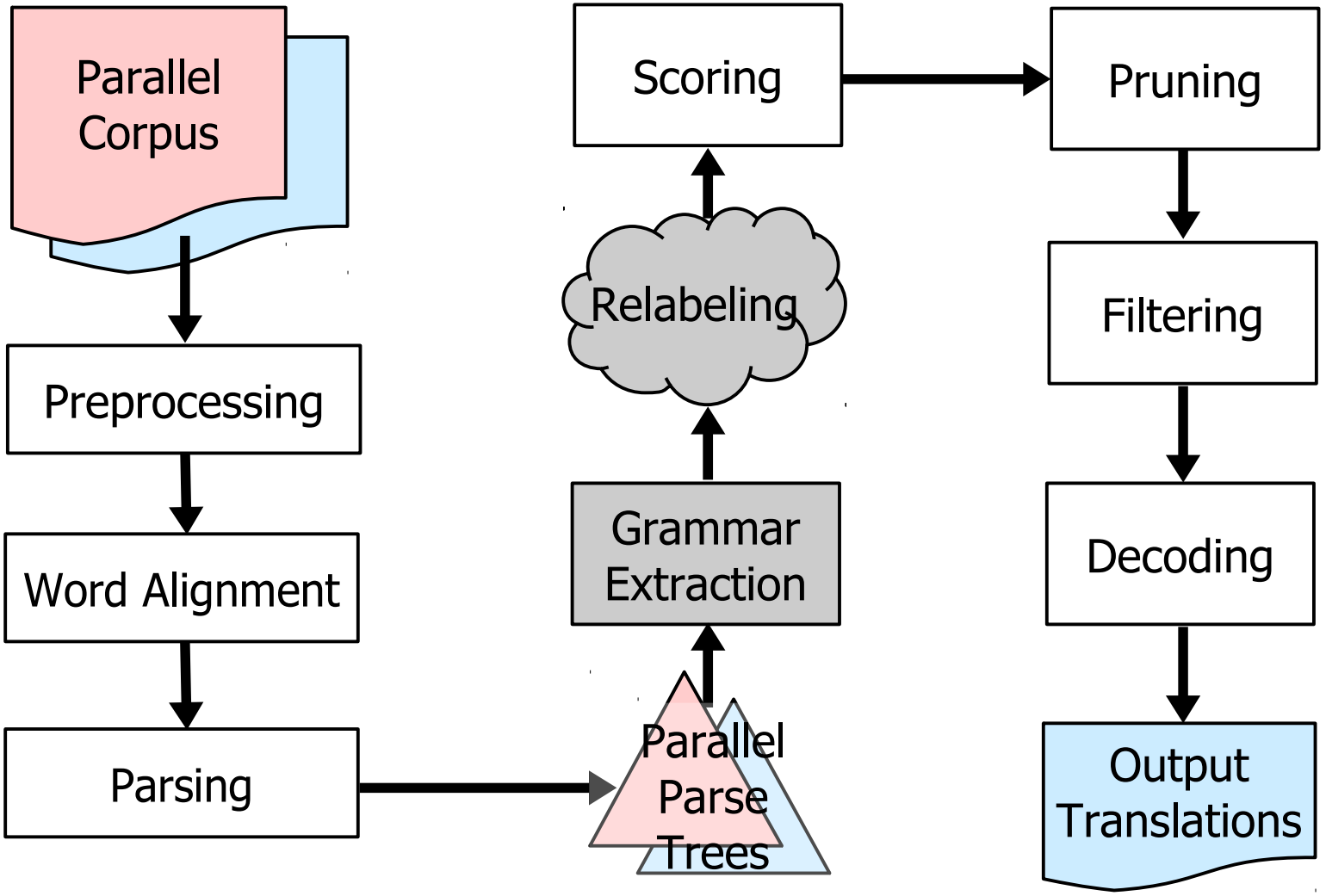


- "Better" will be defined by
  - Significant improvement on automatic metrics
  - Reduced spurious ambiguity and rule sparsity
  - Maintained or improved reordering precision
- Improvements will be demonstrated on at least two language pairs

Outline

- Introduction
- Baseline system and experimental setup
- Label-based system properties
- Extraction and relabeling techniques
- Putting it all together

Experimental Setup



Baseline System

[Hanneman et al. 2010]

- French–English, based on WMT 2010 data
  - 8.6 million sentence pairs of Europarl, news commentary, and UN document text
- Stanford and Berkeley parsers
- Extracted 14 million unique hierarchical rules and 41 million unique phrase pairs
- Pruned to 10,000 most frequent rules and all phrase pairs matching test set
- Joshua SCFG-based decoder

Language Pairs and Data

- French–English
  - Millions of sentence pairs of Europarl, news commentary, UN document, and Web text
  - Annual WMT evaluation
- Arabic–English and Chinese–English
  - Millions of sentence pairs released from LDC
  - Periodic NIST evaluation
- Chinese–English proving ground
  - FBIS corpus of 300,000 sentence pairs

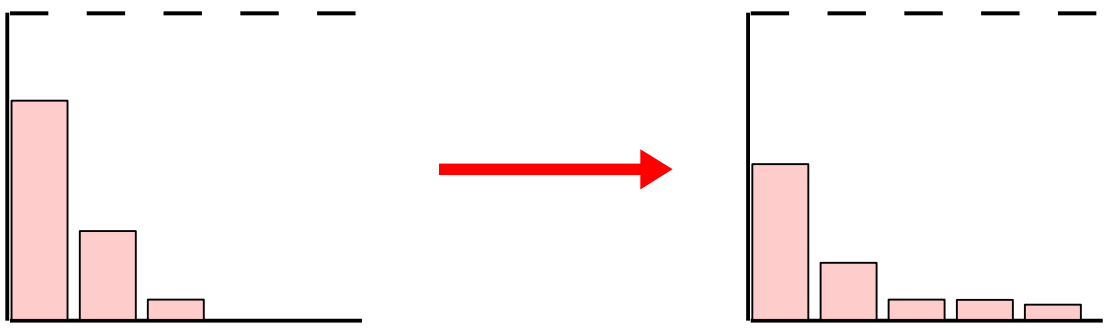
Outline

- Introduction
- Baseline system and experimental setup
- Label-based system properties
  - Spurious ambiguity
  - Rule sparsity
  - Reordering precision
- Extraction and relabeling techniques
- Putting it all together

(1) Spurious Ambiguity

- "Many derivations that are distinct yet have the same model feature vectors and give the same translation" [Chiang 2005]
- Weakens conditional probabilities

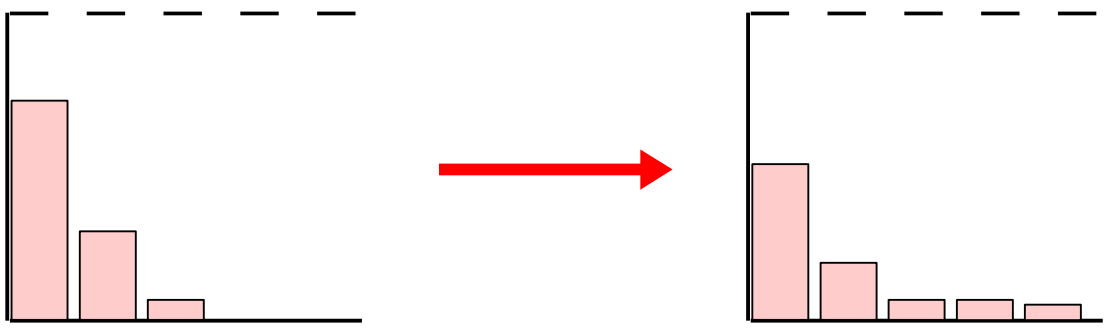
$$\mathrm{P}(\ell_s, \ell_t, r_s \mid r_t) \quad \mathrm{P}(\ell_s, \ell_t \mid r_s) \quad \mathrm{P}(\ell_s, \ell_t \mid r_s, r_t)$$
$$\mathrm{P}(\ell_s, \ell_t, r_t \mid r_s) \quad \mathrm{P}(\ell_s, \ell_t \mid r_t) \quad \mathrm{P}(r_s, r_t \mid \ell)$$



(1) Spurious Ambiguity

- "Many derivations that are distinct yet ~~have the same model feature vectors and~~ give the same translation" [Chiang 2005]
- Weakens conditional probabilities

$$\mathrm{P}(\ell_s, \ell_t, r_s \mid r_t) \quad \mathrm{P}(\ell_s, \ell_t \mid r_s) \quad \mathrm{P}(\ell_s, \ell_t \mid r_s, r_t)$$
$$\mathrm{P}(\ell_s, \ell_t, r_t \mid r_s) \quad \mathrm{P}(\ell_s, \ell_t \mid r_t) \quad \mathrm{P}(r_s, r_t \mid \ell)$$



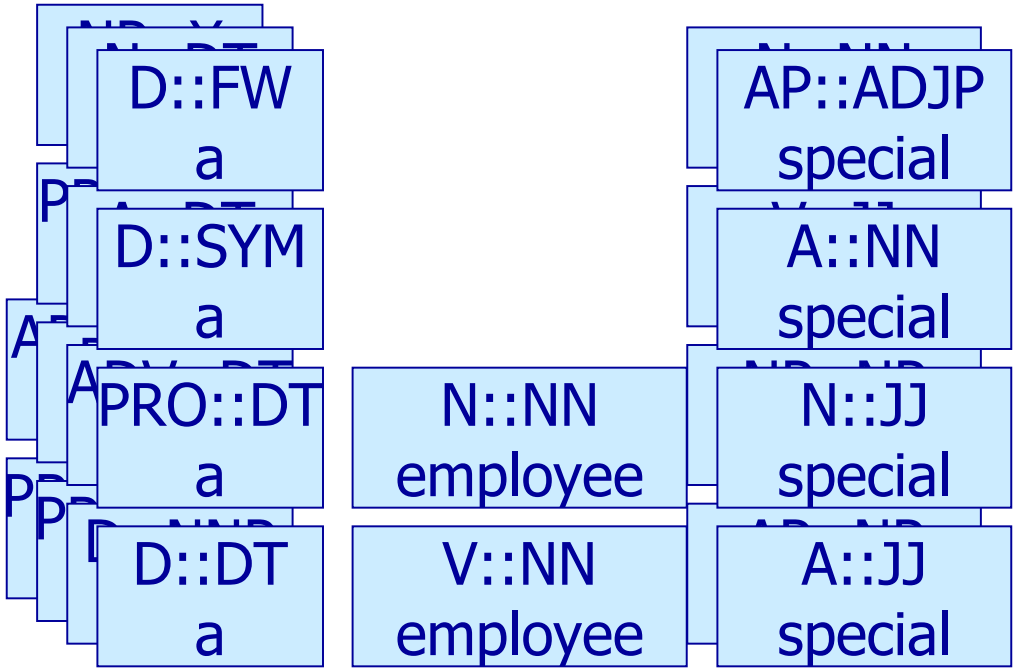
(1) Spurious Ambiguity

- Generates duplicate chart entries with varying left-hand-side labels

une employée spéciale

(1) Spurious Ambiguity

- Generates duplicate chart entries with varying left-hand-side labels



(2) Rule Sparsity

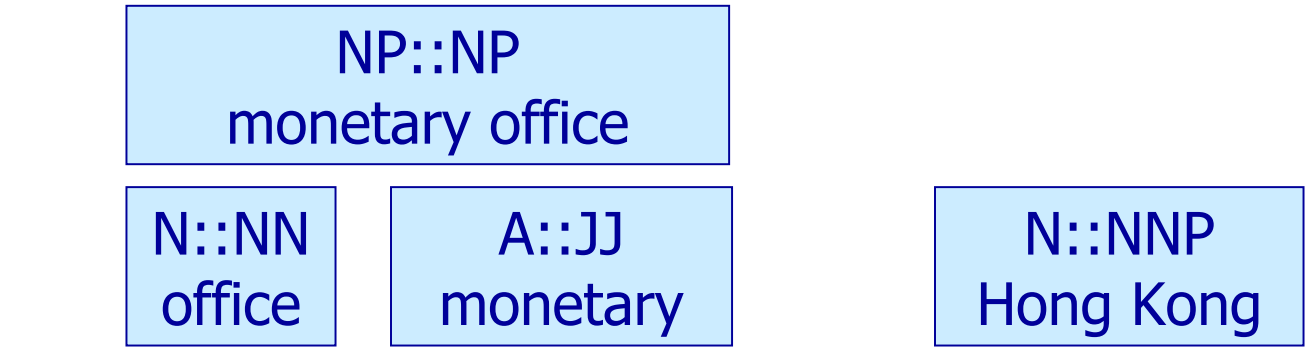
- Necessary rule not found in grammar
- Prevents building of translation fragment

l’ office monétaire de Hongkong

(2) Rule Sparsity

- Necessary rule not found in grammar
- Prevents building of translation fragment

$$\text{NP::NP} \rightarrow [\text{NP}^1 \text{ de } \text{NP}^2]::[\text{NP}^2 \text{ NP}^1]$$
$$\text{NP::NP} \rightarrow [\text{NP}^1 \text{ de } \text{N}^2]::[\text{NN}^2 \text{ NP}^1]$$



(2) Rule Sparsity

- Necessary rule not found in grammar
- Prevents building of translation fragment

$$\text{NP::NP} \rightarrow [\text{NP}^1 \text{ de } \text{NP}^2]::[\text{NP}^2 \text{ NP}^1]$$
$$\text{NP::NP} \rightarrow [\text{NP}^1 \text{ de } \text{N}^2]::[\text{NN}^2 \text{ NP}^1]$$

Need an NP::NP or an N::NN

(3) Reordering Precision

- Restricts rules to only apply in contexts specified by their labels

$$X::X \rightarrow [X^1 \text{ de } X^2]::[X^1\ X^2]$$
$$X::X \rightarrow [X^1 \text{ de } X^2]::[X^2\ X^1]$$

$$PP::X \rightarrow [P^1 \text{ de } NP^2]::[X^1\ X^2]$$
$$NP::X \rightarrow [N^1 \text{ de } N^2]::[X^2\ X^1]$$

$$PP::PP \rightarrow [P^1 \text{ de } NP^2]::[IN^1\ NP^2]$$
$$NP::NP \rightarrow [N^1 \text{ de } N^2]::[NN^2\ NNS^1]$$

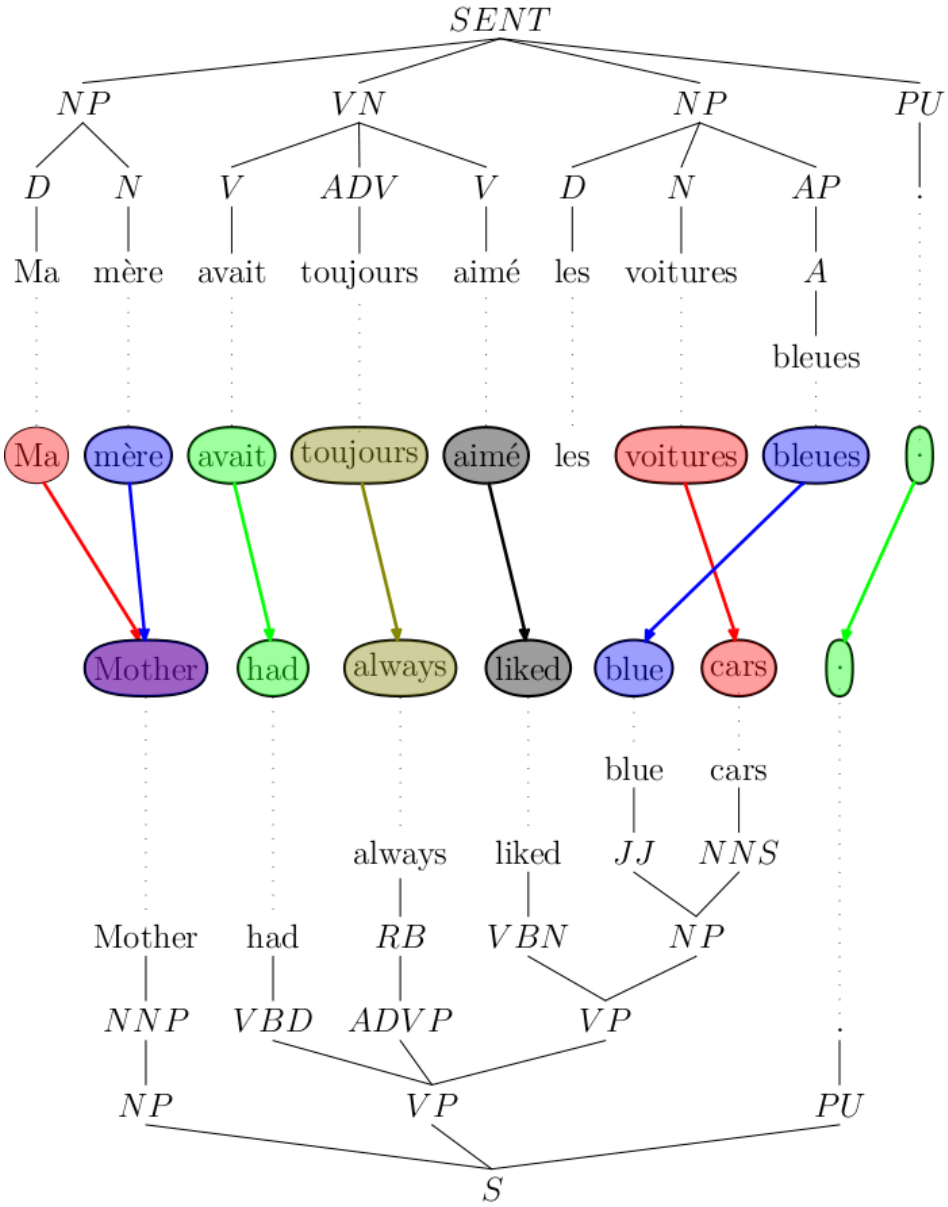
Measuring These Properties

- Simple counting in the grammar
  - Left-hand-side labels per right-hand side
  - Fraction of possible labelings that instantiate a reordering pattern
- Counting within runtime system
  - Cell contents in completed decoder charts
  - Words out of order on a test set
- Constrained decoding [Auli et al. 2009]
  - Reference reachability

Outline

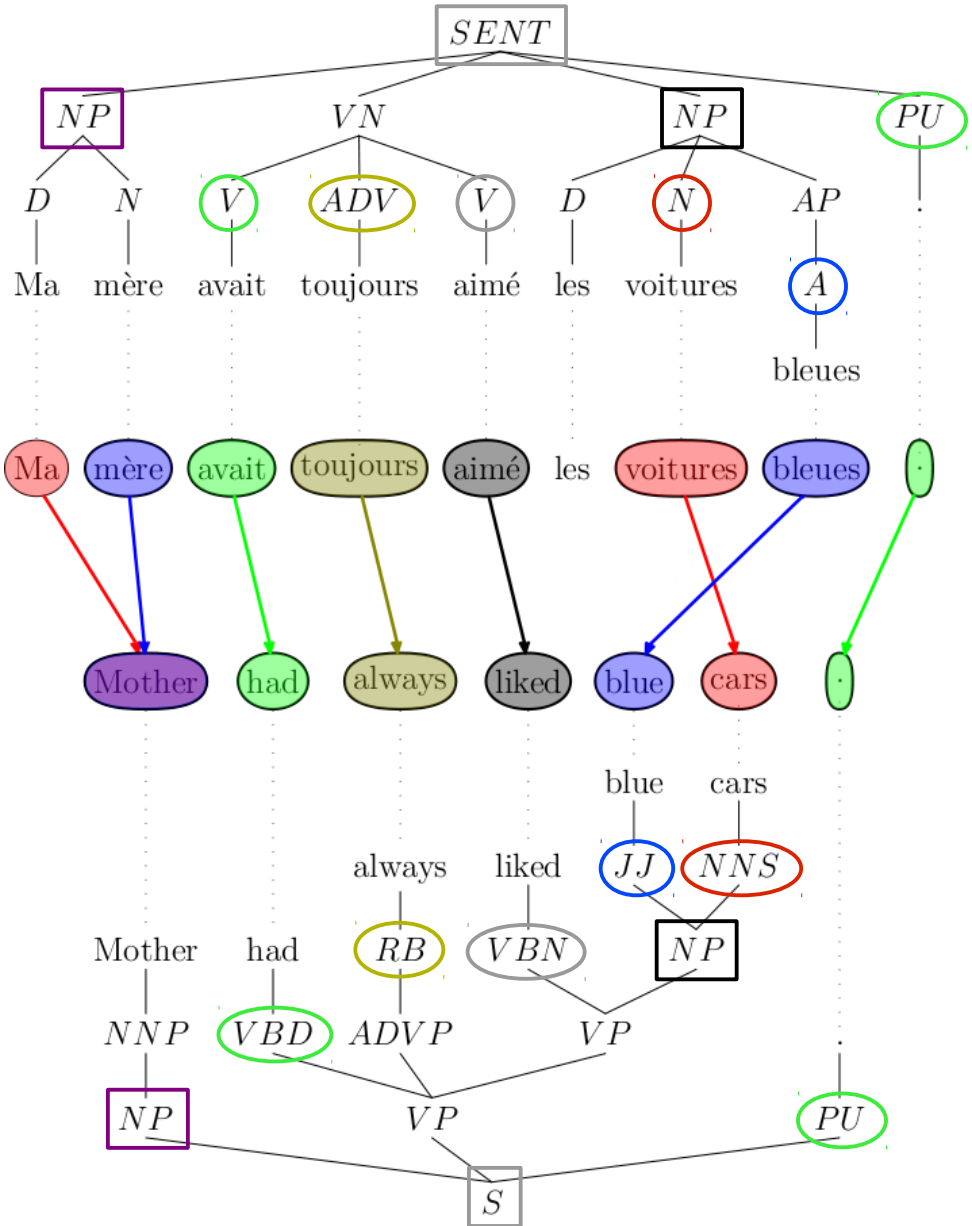
- Introduction
- Baseline system and experimental setup
- Label-based system properties
- Extraction and relabeling techniques
  - General-purpose rule extractor
  - Label collapsing
  - Label refining
  - Correcting local labeling errors
- Putting it all together

(1) General Rule Extractor



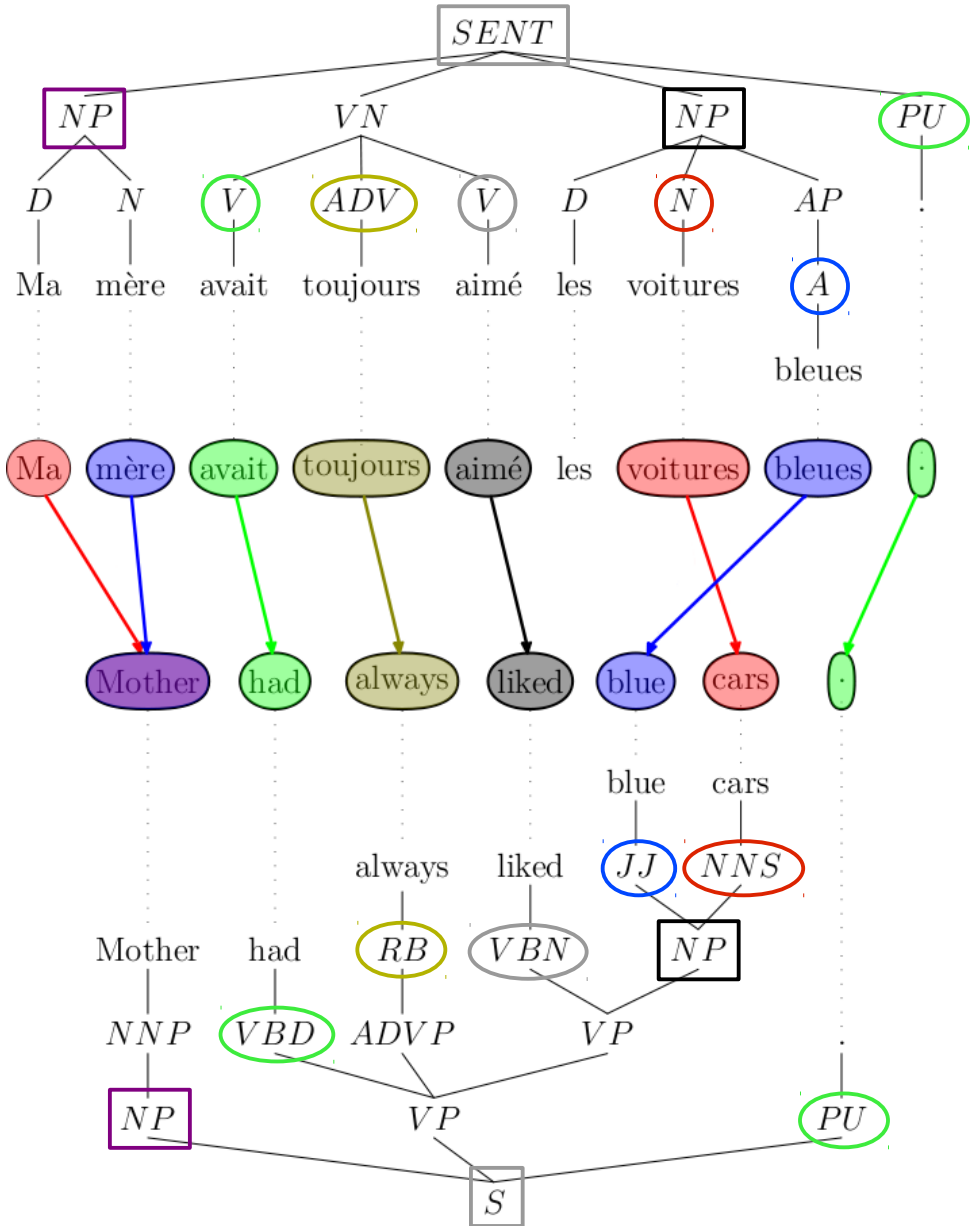
Baseline Extraction Method

[Lavie et al. 2008]



Baseline Extraction Method

[Lavie et al. 2008]



NP::NP → [les voitures bleues]::
[blue cars]

NP::NP → [les N^1 A^2]::
[JJ^2 NNS^1]

Proposed Extensions

- Node alignment
  - Remove constraints on ambiguous alignments
  - Align single node to sequence of sibling nodes
- Rule extraction
  - Produce multiple tree decompositions

$$\text{NP::NP} \rightarrow \text{[les voitures bleues]::[blue cars]}$$

$$\text{NP::NP} \rightarrow \text{[les voitures A}^{2}\text{]::[JJ}^{2}\text{ cars]}$$

$$\text{NP::NP} \rightarrow \text{[les N}^{1}\text{ bleues]::[blue NNS}^{1}\text{]}$$

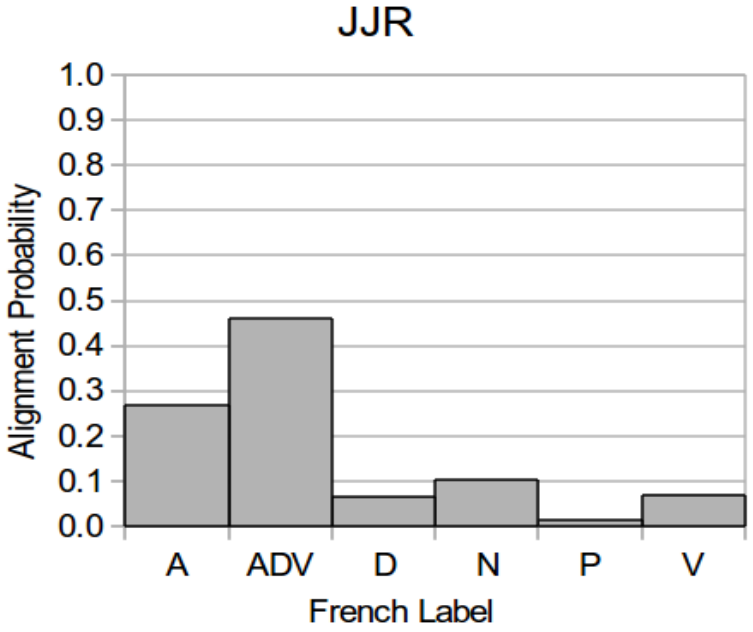
$$\text{NP::NP} \rightarrow \text{[les N}^{1}\text{ A}^{2}\text{]::[JJ}^{2}\text{ NNS}^{1}\text{]}$$

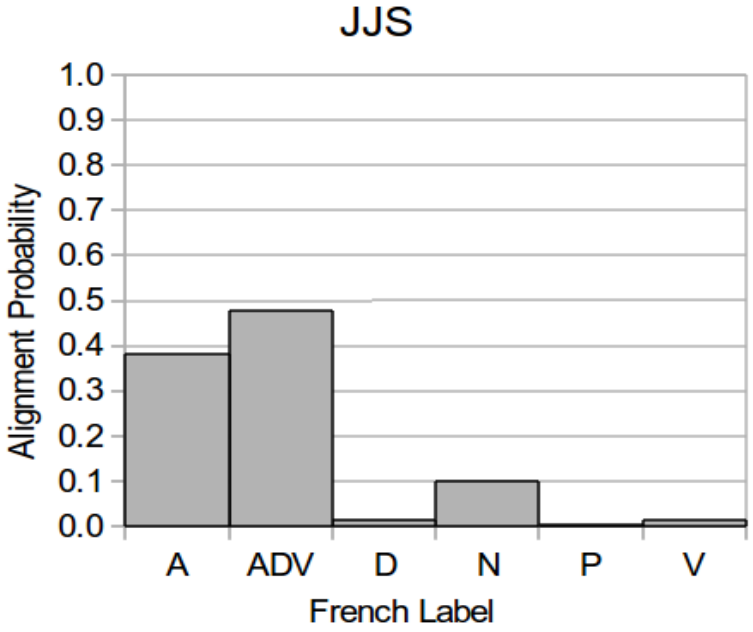
(2) Label Collapsing

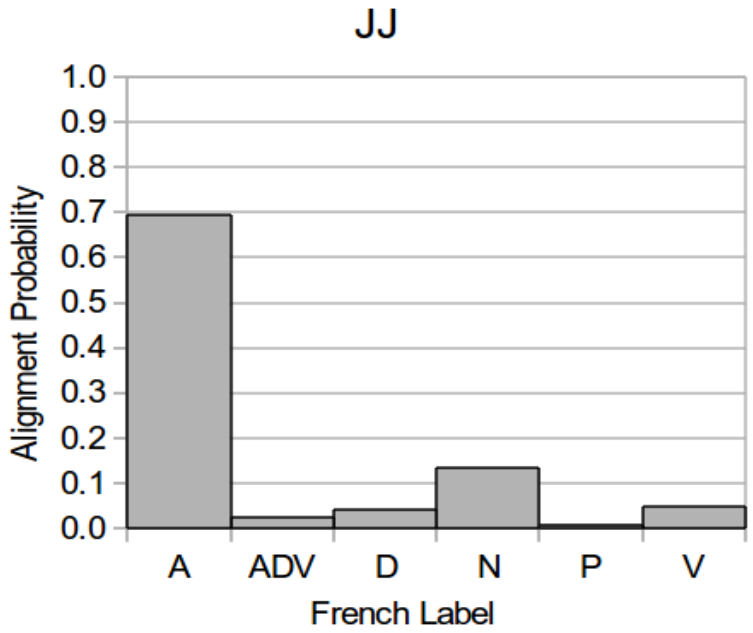
- Spurious ambiguity and rule sparsity caused by large label sets
- Idea: Cluster and collapse the label set
- Intuition: Categories that translate differently still "deserve" different labels
  - English JJ vs. JJR and JJS

the large car	la grande voiture
the larger car	la plus grande voiture
the largest car	la voiture la plus grande

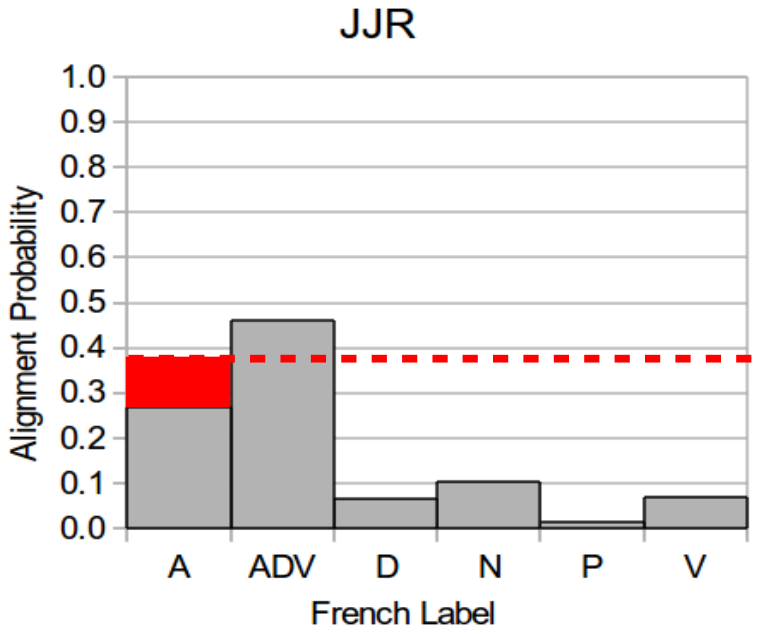
Alignment Distribution Difference

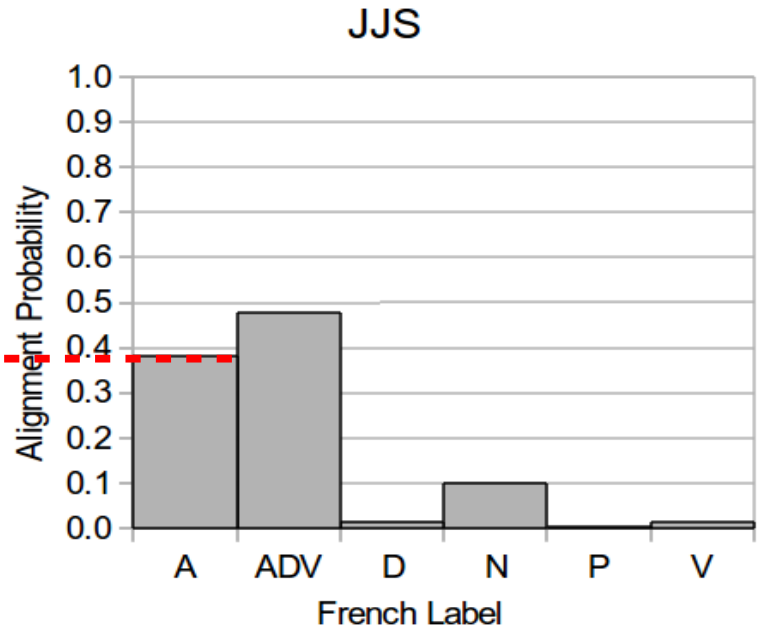


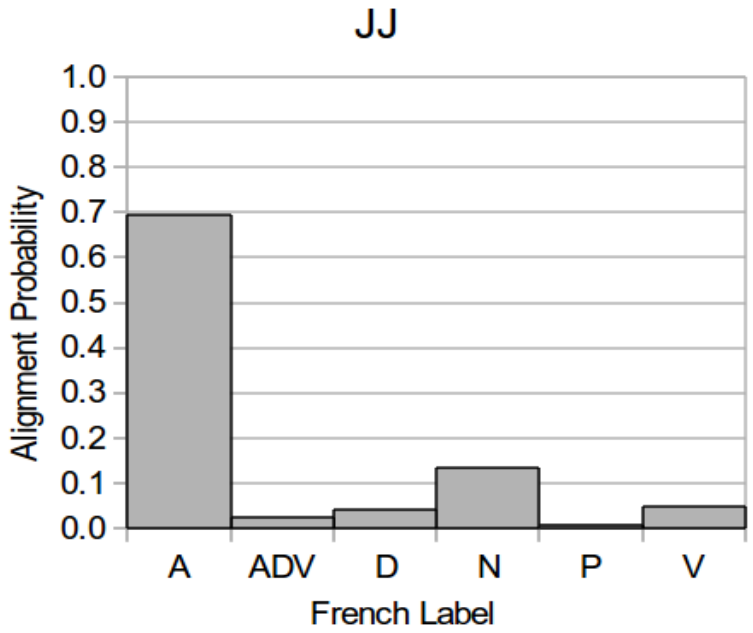




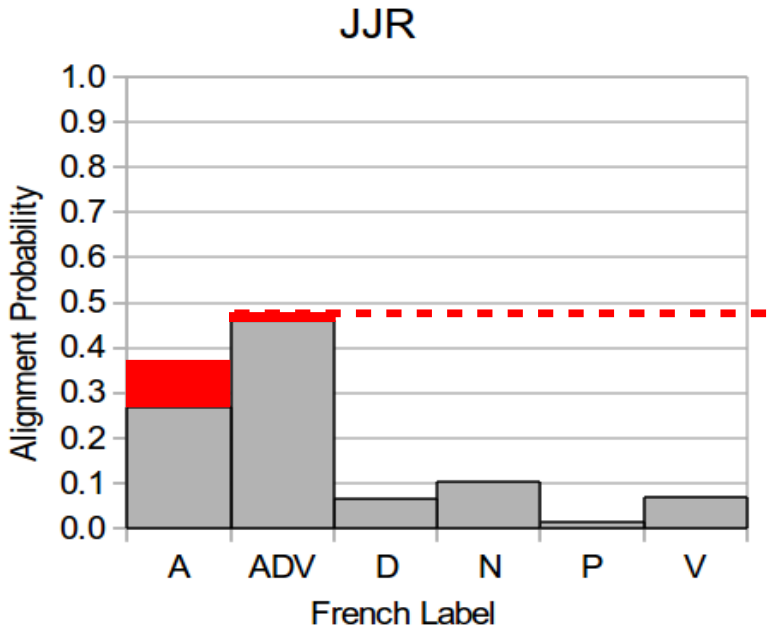
Alignment Distribution Difference

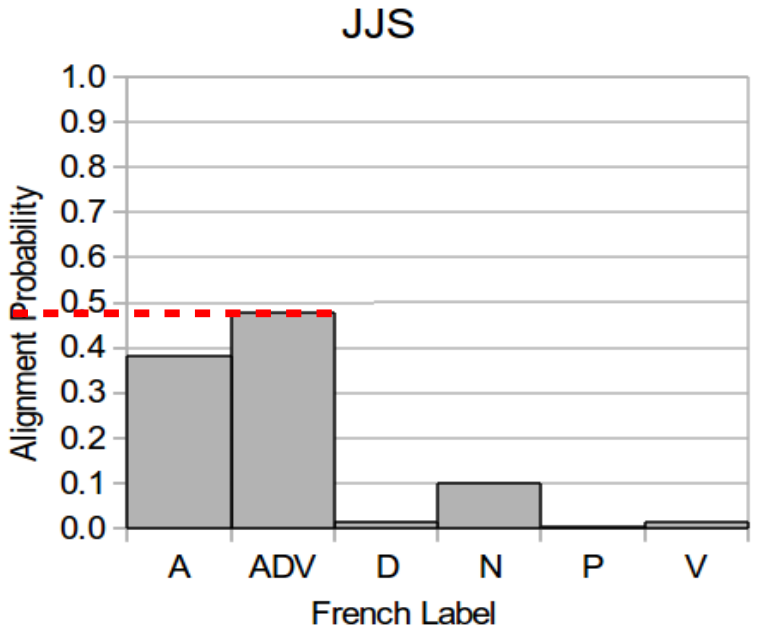


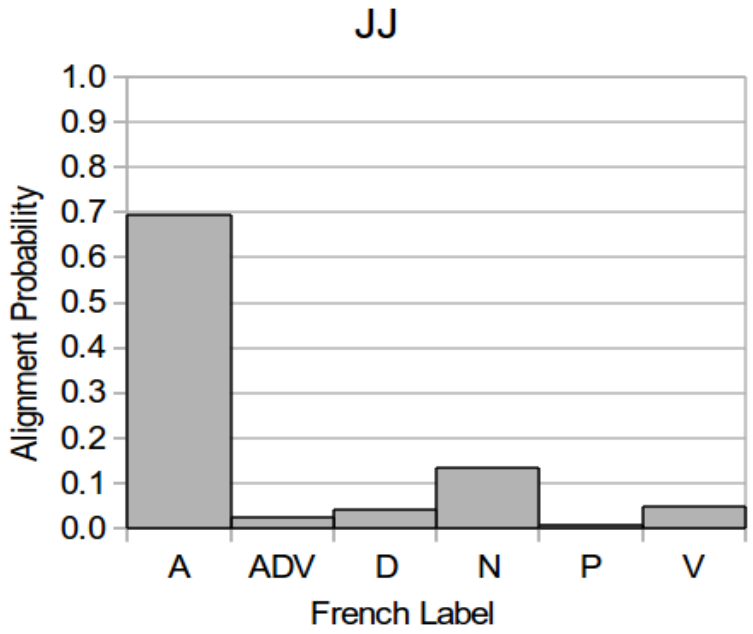




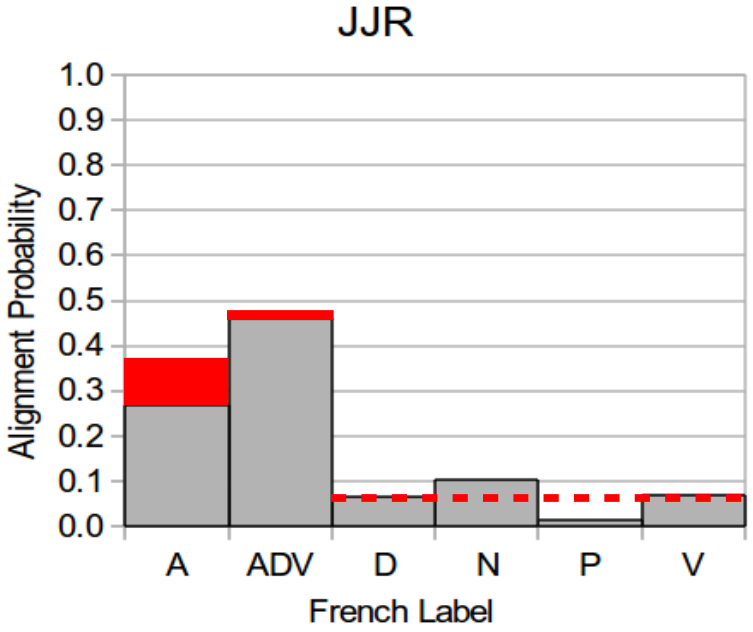
Alignment Distribution Difference

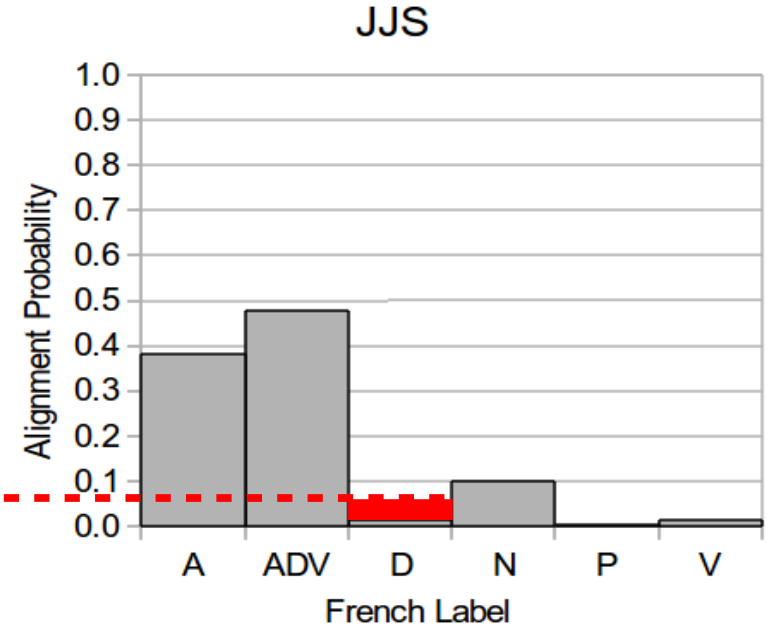


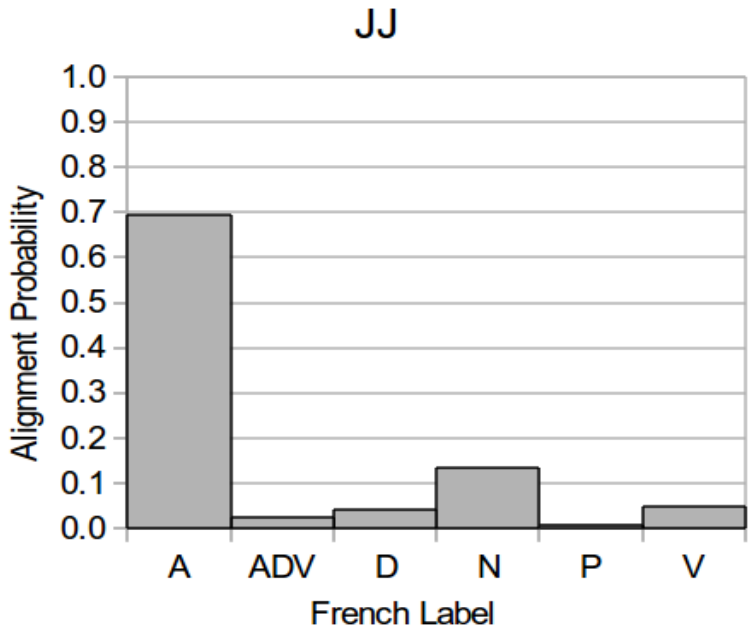




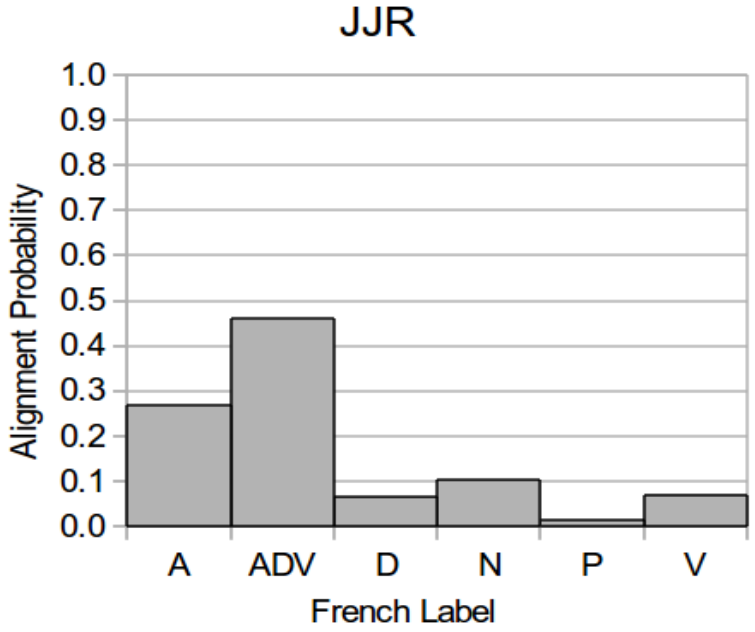
Alignment Distribution Difference



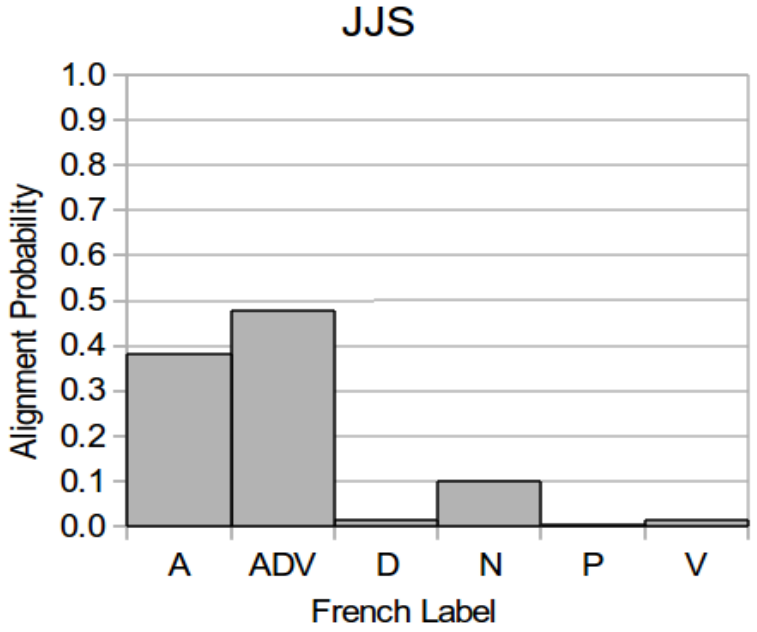




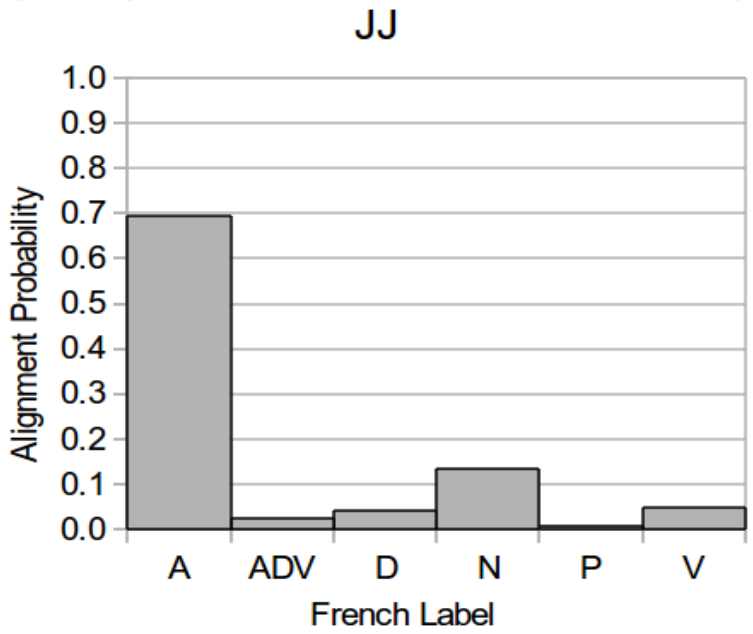
Alignment Distribution Difference



0.2688



0.9952

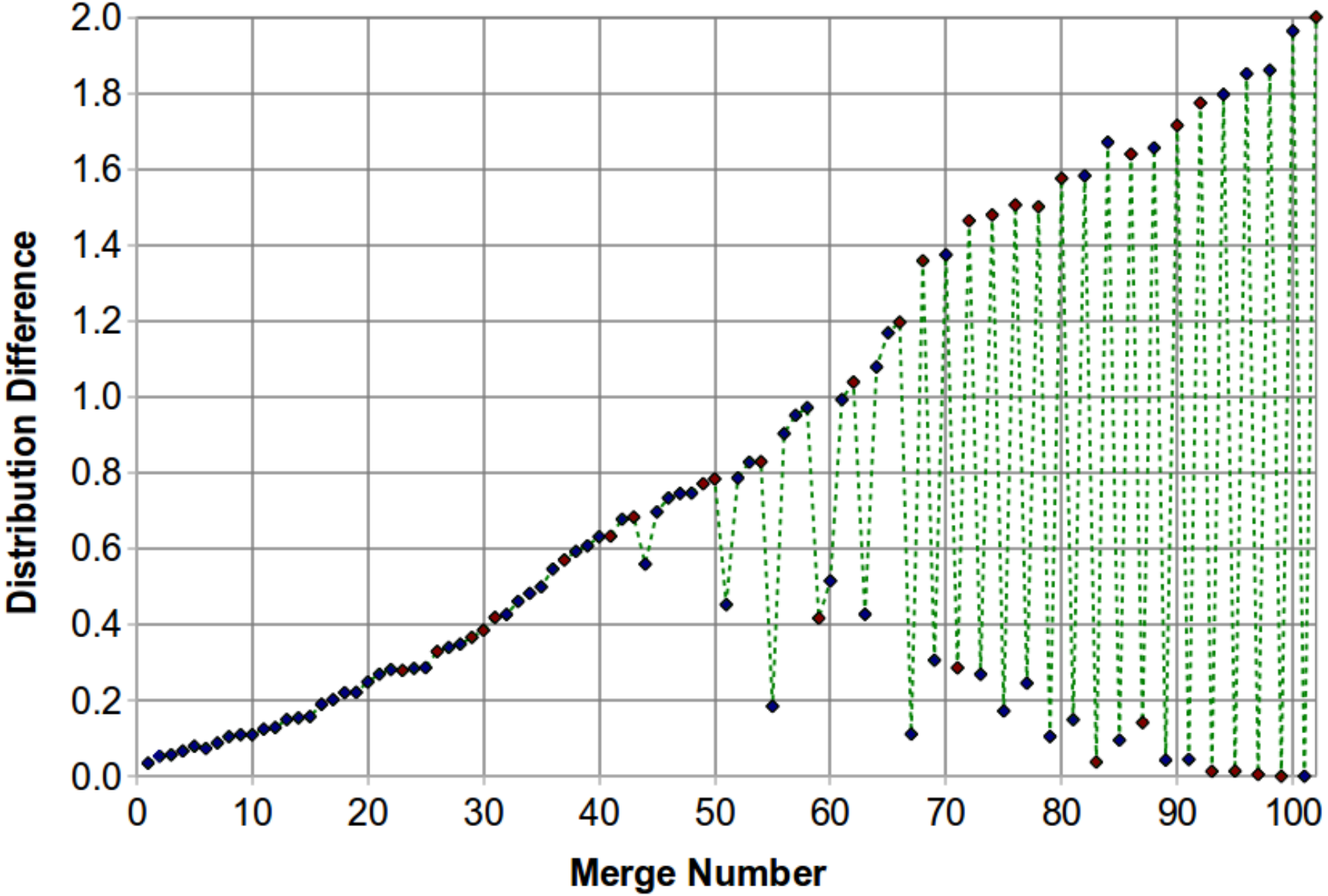


0.9114

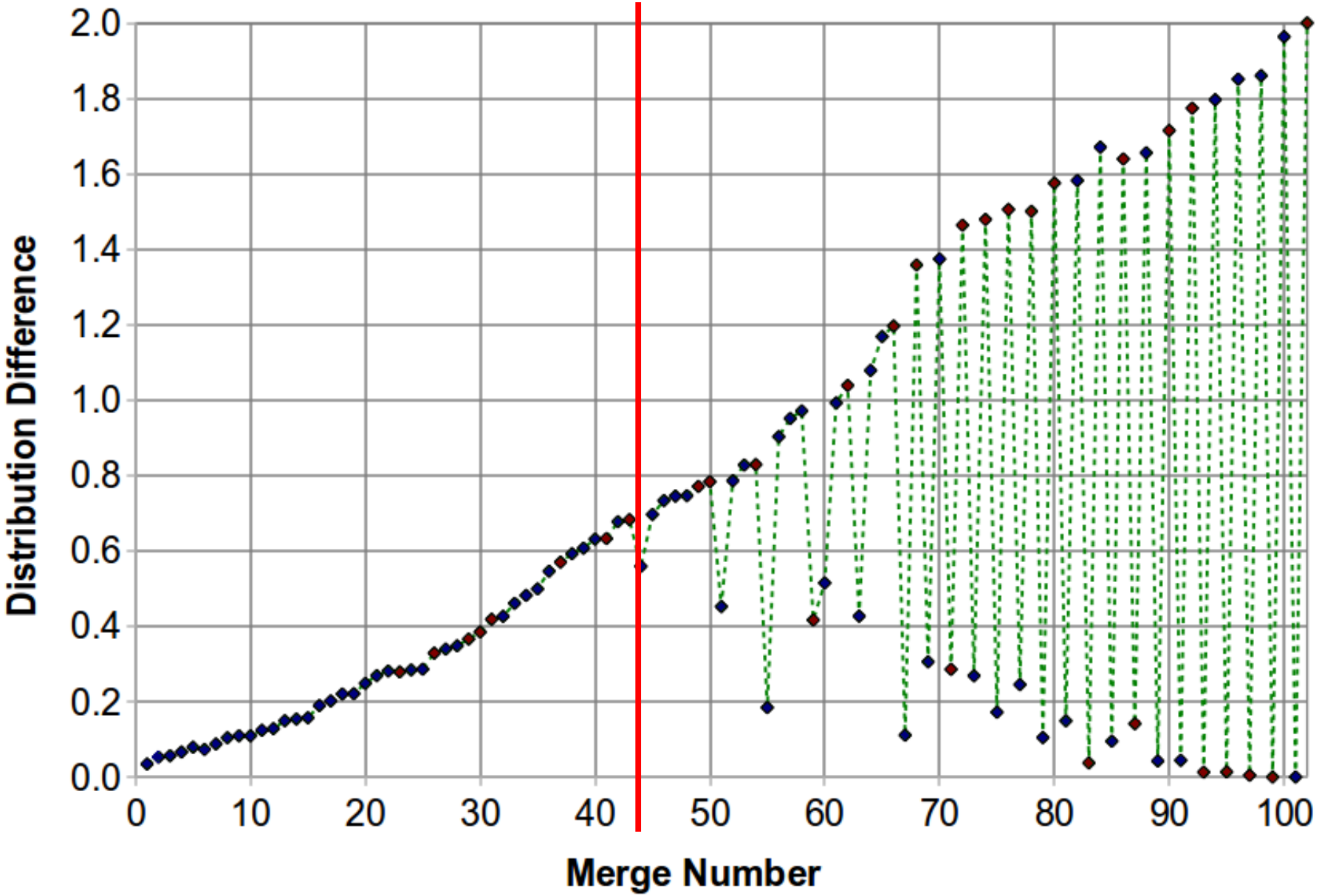
Label Collapsing Algorithm

- Greedy procedure
  - Compute distribution difference between all pairs of source and target labels
  - Merge the label pair with smallest value
  - Update alignment statistics
  - Repeat
- Stopping criterion...?

Label Collapsing Algorithm



Label Collapsing Algorithm



Results: Grammar

- Label set
  - Before: 33 French × 72 English ⇒ 1134 joint
  - After: 25 French × 37 English ⇒ 502 joint
- Spurious ambiguity
  - 1000 most frequent phrase pairs have 21% fewer left-hand-side labels
- Rule sparsity
  - Label sequences 7 times more likely to fit 10 frequent reordering patterns
  - More rule applications in MT test sets

Results: MT Output

- Collapsing the label set improves scores
  - news-test2009

	BLEU	METR	TER
Original	21.75	51.96	59.44
Collapsed	22.78	53.00	59.16

  - news-test2010

	BLEU	METR	TER
Original	22.03	53.13	58.00
Collapsed	23.14	54.03	57.86

Proposed Work

- Investigate choice of
  - Stopping criterion
  - Distance metric
  - Greedy collapsing vs. all at once
- Run collapsing from different initial label granularities [Petrov et al. 2006]
  - Default: S, NP, NNS, A, VPinf
  - Refined: S-1, S-4, NP-2, NP-17, NNS-8, A-11, A-13, VPinf-3, etc.

(3) Label Refining

- Bilingual SCFG rules encode knowledge not used in monolingual parsers
- Idea: Split existing labels using reordering information
- Intuition: Rules that differ only in target-side order indicate hidden categories

$$\text{NP::NP} \rightarrow [\text{DT}^1\ \text{JJ}^2\ \text{NN}^3]::[\text{D}^1\ \text{N}^3\ \text{A}^2]$$

$$\text{NP::NP} \rightarrow [\text{DT}^1\ \text{JJ}^2\ \text{NN}^3]::[\text{D}^1\ \text{A}^2\ \text{N}^3]$$

Proposed Algorithm

- Identify rule set of interest

$$\text{NP::NP} \rightarrow [\text{DT}^1\ \text{JJ}^2\ \text{NN}^3]::[\text{D}^1\ \text{N}^3\ \text{A}^2]$$

$$\text{NP::NP} \rightarrow [\text{DT}^1\ \text{JJ}^2\ \text{NN}^3]::[\text{D}^1\ \text{A}^2\ \text{N}^3]$$

- From parsed corpus, tabulate what plugs into the rules at each extracted instance
- Items much more likely to plug into one rule than other form a category subtype

Refining Example

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[\ \_\_\_\_\_\ ]$$

$D^1\ N^3\ A^2$	
un	0.3165
une	0.2654
la	0.1288
le	0.0815
cette	0.0428
l’	0.0371
ce	0.0368
…	…

$D^1\ A^2\ N^3$	
une	0.2545
un	0.2067
la	0.1688
le	0.1054
cette	0.0563
Le	0.0515
La	0.0435
…	…

Refining Example

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[\ \_\_\_\_\_\ ]$$

$D^1\ N^3\ A^2$			$D^1\ A^2\ N^3$	
un	0.3165	0.4154	une	0.2545
une	0.2654	⟷	un	0.2067
la	0.1288		la	0.1688
le	0.0815		le	0.1054
cette	0.0428		cette	0.0563
l’	0.0371		Le	0.0515
ce	0.0368		La	0.0435
…	…		…	…

Refining Example

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[\ \_\_\_\_\_\ ]$$

$D^1\ \mathbf{N^3}\ A^2$		$D^1\ A^2\ \mathbf{N^3}$	
rôle	0.0349	fois	0.0397
position	0.0241	temps	0.0355
question	0.0213	point	0.0258
communauté	0.0212	majorité	0.0212
solution	0.0172	question	0.0199
situation	0.0171	chose	0.0193
base	0.0154	problème	0.0174
…	…	…	…

Refining Example

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[\ \_\_\_\_\_\ ]$$

$D^1\ N^3\ A^2$		$D^1\ A^2\ N^3$	
commune	0.0338	même	0.0957
important	0.0286	nouvelle	0.0837
politique	0.0285	première	0.0676
internationale	0.0267	bonne	0.0593
européenne	0.0218	nouveau	0.0521
européen	0.0186	deuxième	0.0509
juridique	0.0144	bon	0.0406
…	…	…	…

Refining Example

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[D^1\ N^3\ \mathbf{AA^2}]$$

$$NP::NP \rightarrow [DT^1\ JJ^2\ NN^3]::[D^1\ \mathbf{AB^2}\ N^3]$$

AA: commune, politique, internationale, européenne, européen, juridique, ...

AA & AB: important, importante, ...

AB: même, nouvelle, première, bonne, nouveau, deuxième, bon, autre, ...

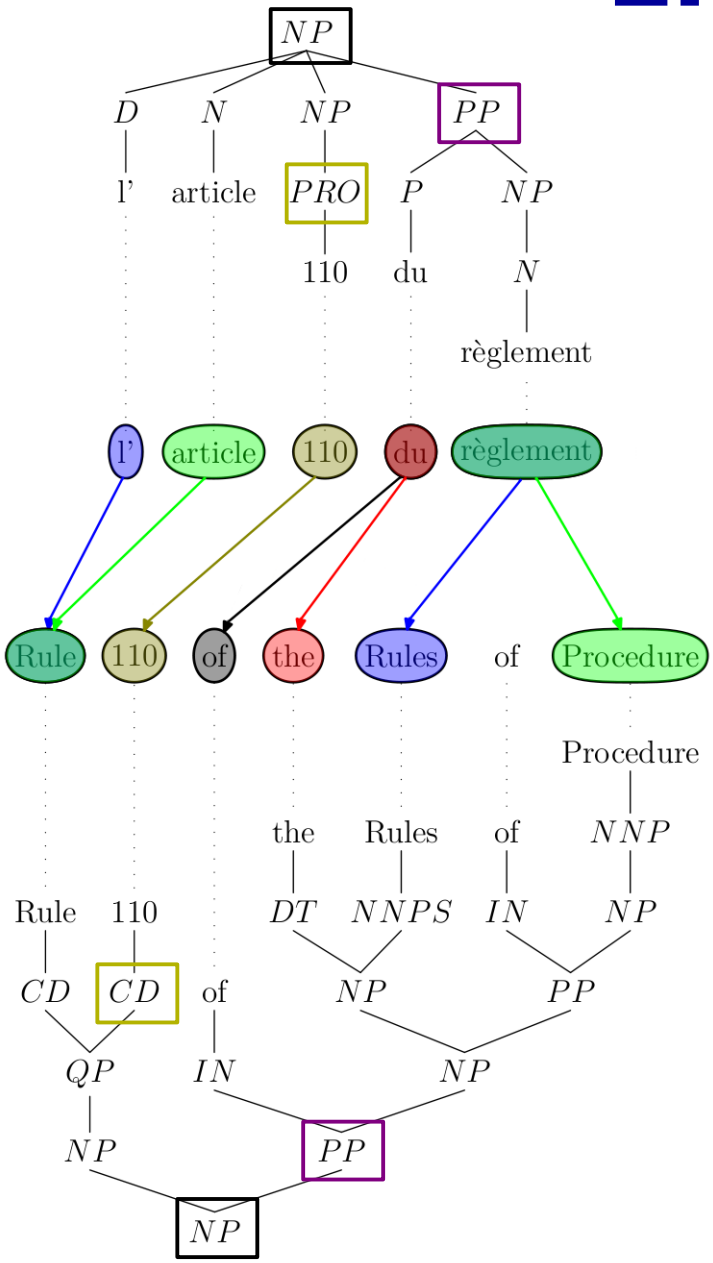
(4) Correcting Local Label Errors

- Sentences parsed independently may contain local labeling errors
- Idea: Correct them using rules extracted from the entire parsed parallel corpus
- Intuition: Rule probabilities from whole corpus can fix label errors; fixing local label errors improves rule probabilities

Proposed Algorithm

- EM algorithm
  - Initial grammar extraction: compute P(LHS | RHS) for each rule
  - E step: Re-derive most likely labeled version of each aligned tree pair using rule scores
  - M step: Re-extract grammar from modified trees and re-compute rule scores
  - Repeat E and M steps until grammar doesn't change

EM Example

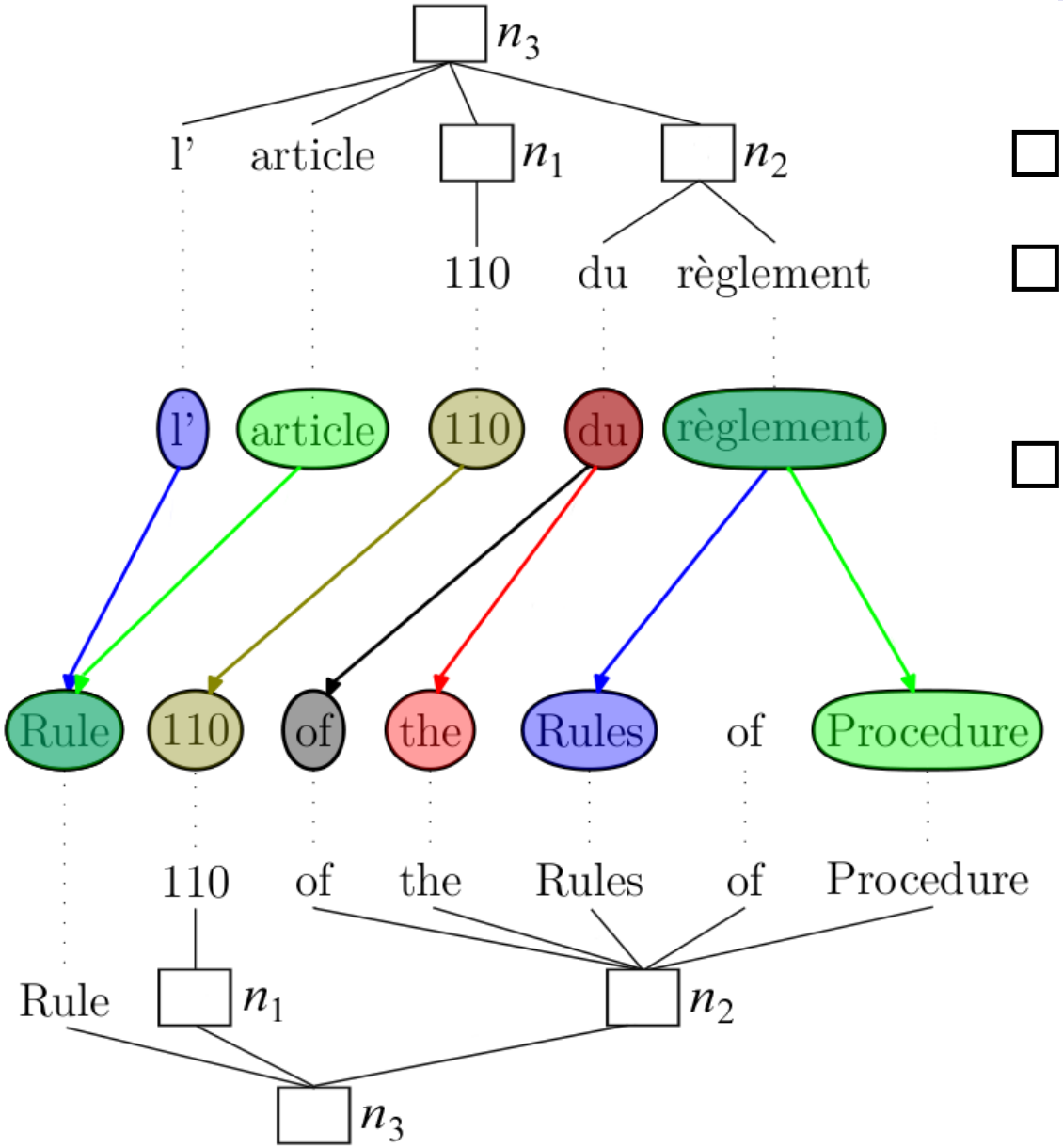


PRO::CD → [110]::[110]

PP::PP → [du règlement]::
[of the Rules of Procedure]

NP::NP → [l' article **PRO**^{1} PP^{2}]::
[Rule CD^{1} PP^{2}]

EM Example

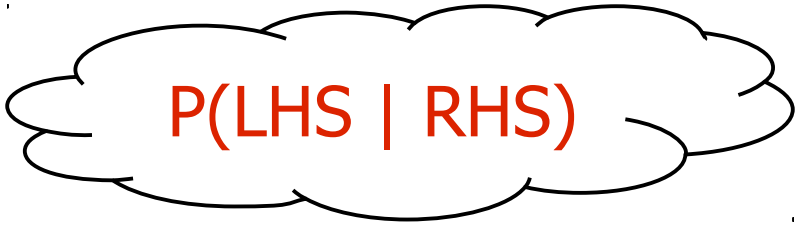


$\square::\square \rightarrow$ [110]::[110]

$\square::\square \rightarrow$ [du règlement]::
[of the R. of P.]

$\square::\square \rightarrow$ [l' article $\square^1$ $\square^2$]::
[Rule $\square^1$ $\square^2$]

EM Example



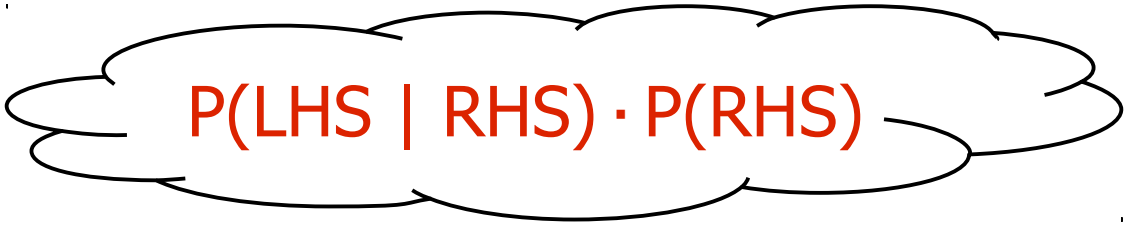
n_1: □::□ → [110]::[110]

D::CD	0.5081
PRO::CD	0.2789
N::CD	0.1004
…	…

n_2: □::□ → [du règlement]::[of the Rules of Procedure]

PP::PP	0.7736
NP::PP	0.2264

EM Example



n_3: $\square::\square \rightarrow [\text{l' article } \square^1\ \square^2]::[\text{Rule } \square^1\ \square^2]$

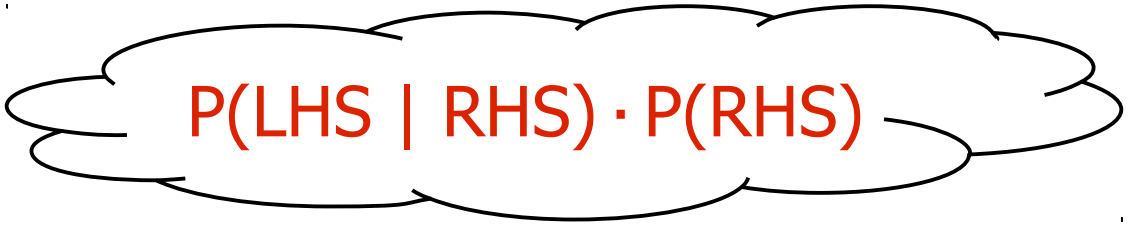
NP::NP → [l' article PRO^1 PP^2]::[Rule CD^1 PP^2]

1 · 0.2789 · 0.7736 = 0.2158

D::CD	0.5081
PRO::CD	0.2789
N::CD	0.1004
…	…

PP::PP	0.7736
NP::PP	0.2264

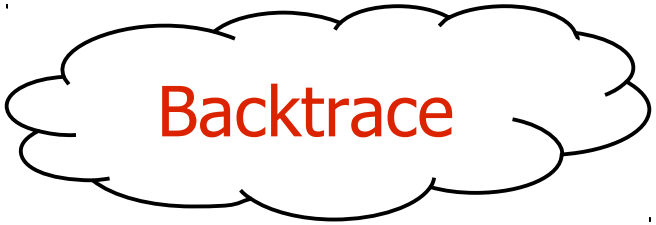
EM Example



n_3: $\square::\square \rightarrow$ [l' article $\square^1$ $\square^2$]::[Rule $\square^1$ $\square^2$]

NP::NP 0.7018
NP::VP 0.0023

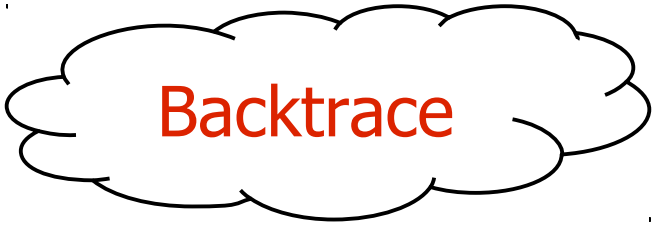
EM Example



n_3: NP::NP → [l' article □[1] □[2]]::[Rule □[1] □[2]]

NP::NP	0.7018
NP::VP	0.0023

EM Example

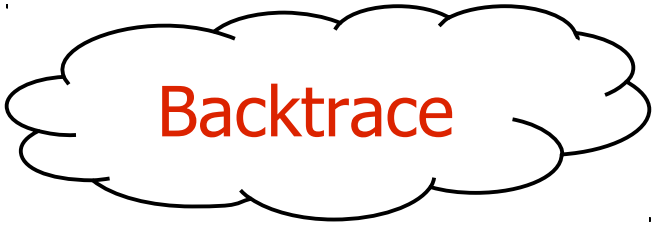


n_3: NP::NP → [l' article D^1 PP^2]::[Rule CD^1 PP^2]

NP::NP 0.7018
NP::VP 0.0023

NP::NP → [l' article D^1 PP^2]::[Rule CD^1 PP^2]
$1 \cdot 0.5081 \cdot 0.7736 = 0.3931$

EM Example



n_3: NP::NP → [l' article D^1 PP^2]::[Rule CD^1 PP^2]

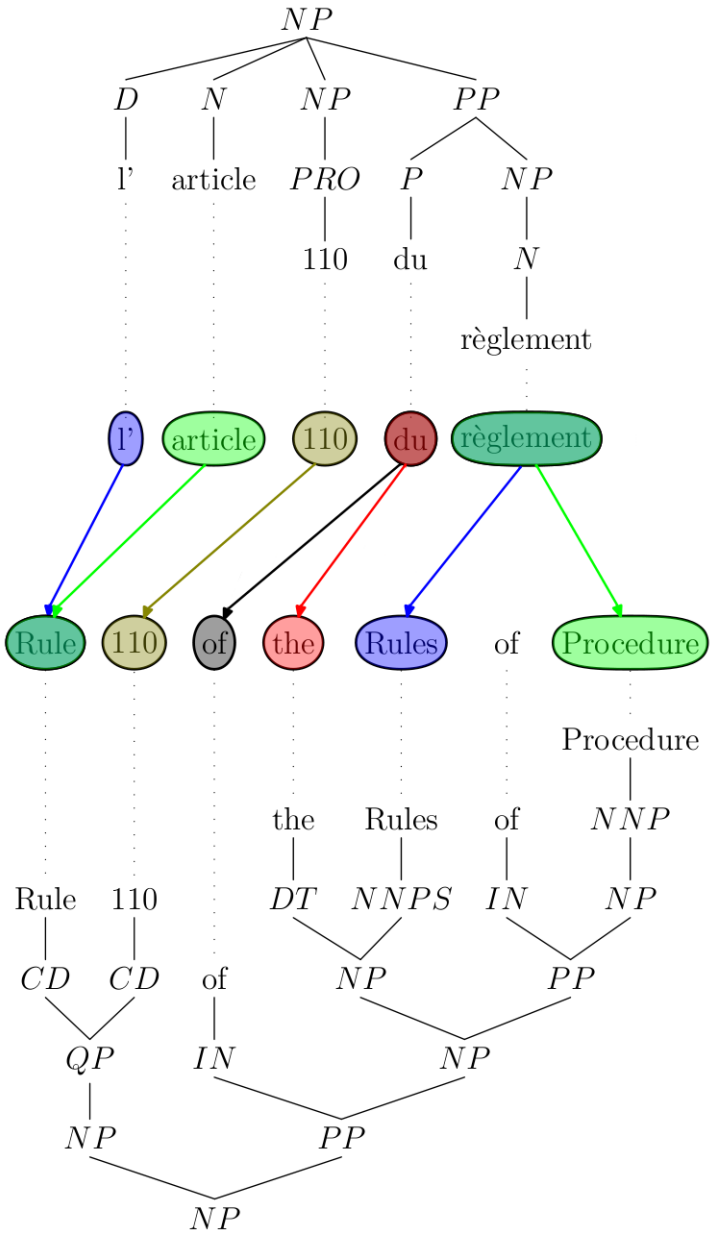
NP::NP 0.7018
NP::VP 0.0023

NP::NP → [l' article D^1 PP^2]::[Rule CD^1 PP^2]
$1 \cdot 0.5081 \cdot 0.7736 = 0.3931$

n_1: D::CD → [110]::[110]

n_2: PP::PP → [du règlement]::[of the Rules of Procedure]

EM Example



D::CD → [110]::[110]

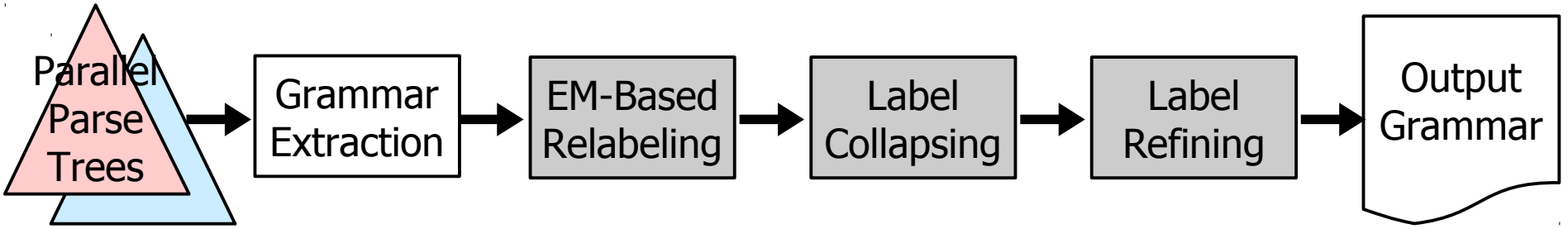
PP::PP → [du règlement]::
[of the Rules of Procedure]

NP::NP → [l' article **D**^{1} PP^{2}]::
[Rule CD^{1} PP^{2}]

Outline

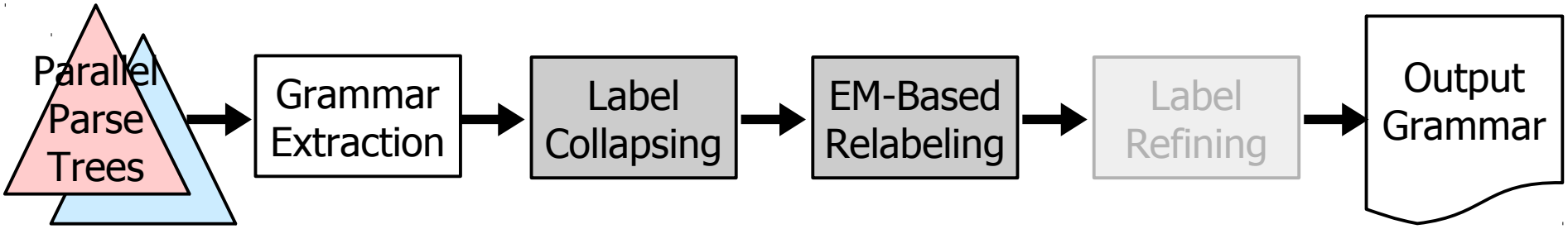
- Introduction
- Baseline system and experimental setup
- Label-based system properties
- Extraction and relabeling techniques
- Putting it all together

Proposed Combination: Default



- Large-scale system with standard labels
  - Remove easy-to-fix occasional labeling errors first (–SA)
  - Remove redundant labels next (–SA, –RS)
  - Add necessary latent categories (+RP)

Proposed Combination: Variants



- Fine-grained input categories
  - Move label collapsing to first (–fragmentation)
  - Remove label refining (focus on RS)
- Small-data systems
  - Remove label refining (focus on RS)

Work Summary

	Completed Work	Proposed Work
Spurious ambiguity, rule sparsity, reordering precision	Defined, measured by simple counting	Measured within system, constrained decoding
General-purpose grammar extractor	Node alignment stage	Rule extraction stage
Label collapsing	Basic algorithm, end-to-end experiments	Algorithm extensions
Label refining	Motivational example, basic approach	Scoping, end-to-end experiments
EM-based relabeling	Motivational example, basic approach	Implementation with grammar extractor, end-to-end experiments
Combined relabeling scheme	Default proposals for different systems	Experimentation

Timeline

Time Period	Work
Jan.-March 2011	Implementation of grammar extractor Full label collapsing experiments Tests of extensions to label collapsing algorithm Submission of EMNLP paper (March 23)
April-July 2011	Development of constrained decoding setup Development of label refining techniques Full label refining experiments French–English submission to WMT 2011
Aug.-Oct. 2011	Development of EM-based relabeling framework Full EM-based relabeling experiments Arabic or Chinese submission to NIST 2011
Nov.-Dec. 2011	Finalization of all individual-technique experiments
Jan.-March 2012	Development of combined-technique pipeline Finalization of systems and experiments
April-May 2012	Thesis writing Thesis defense

Grammar Pruning

- Hiero
  - Phrase pairs limited to length 10
  - Maximum two non-adjacent nonterminals in rules
  - No fully abstract rules
- GHKM
  - Phrase pairs filtered to sentence level
  - Save *n* most frequent translations per source RHS, or all rules extracted at least *k* times (*n* = 40, *k* = 50)
- Our systems
  - Phrase pairs filtered to test set level
  - Save *n* most globally frequent rules (*n* = 10,000)

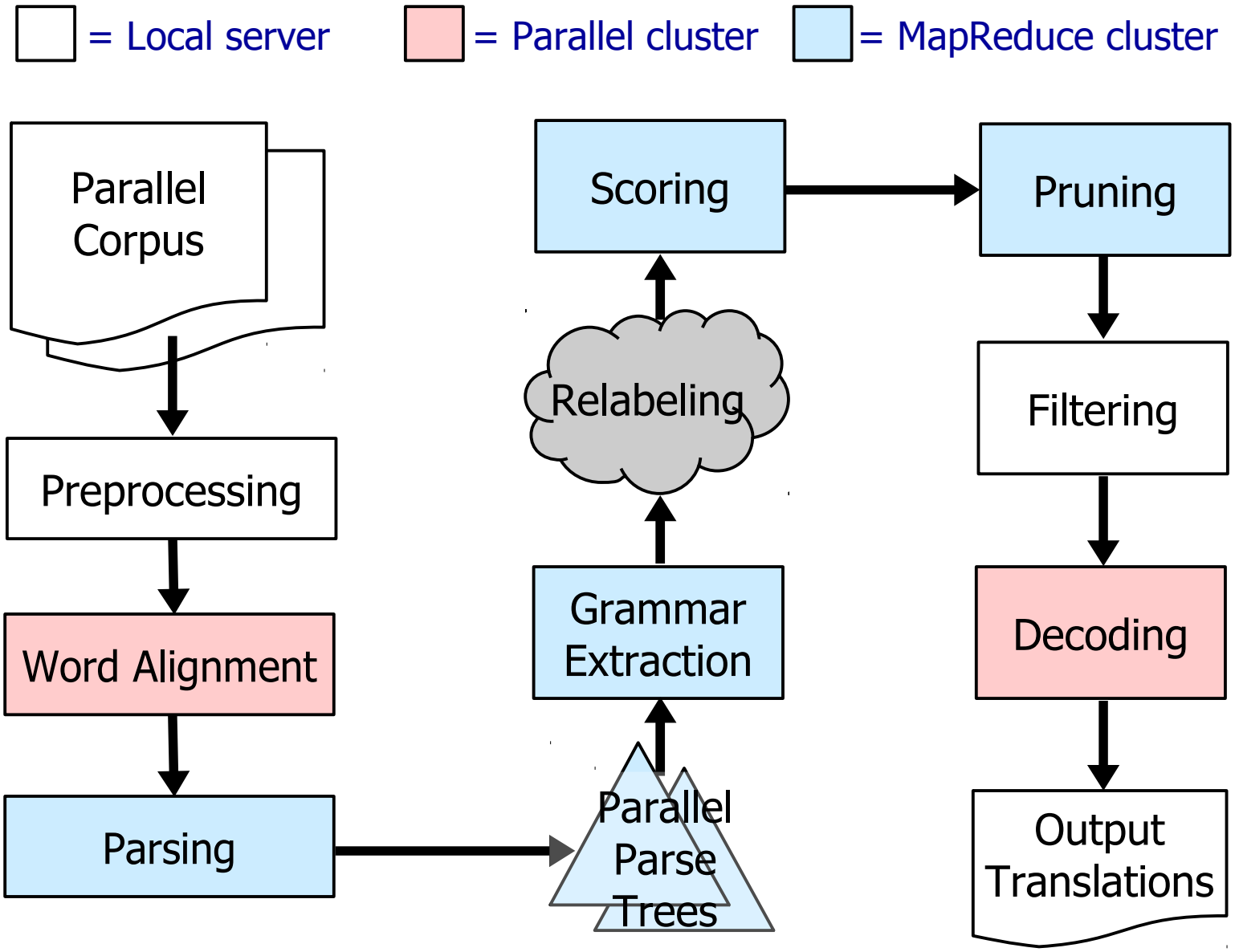
The State of the Art

- WMT news-test2010 test set

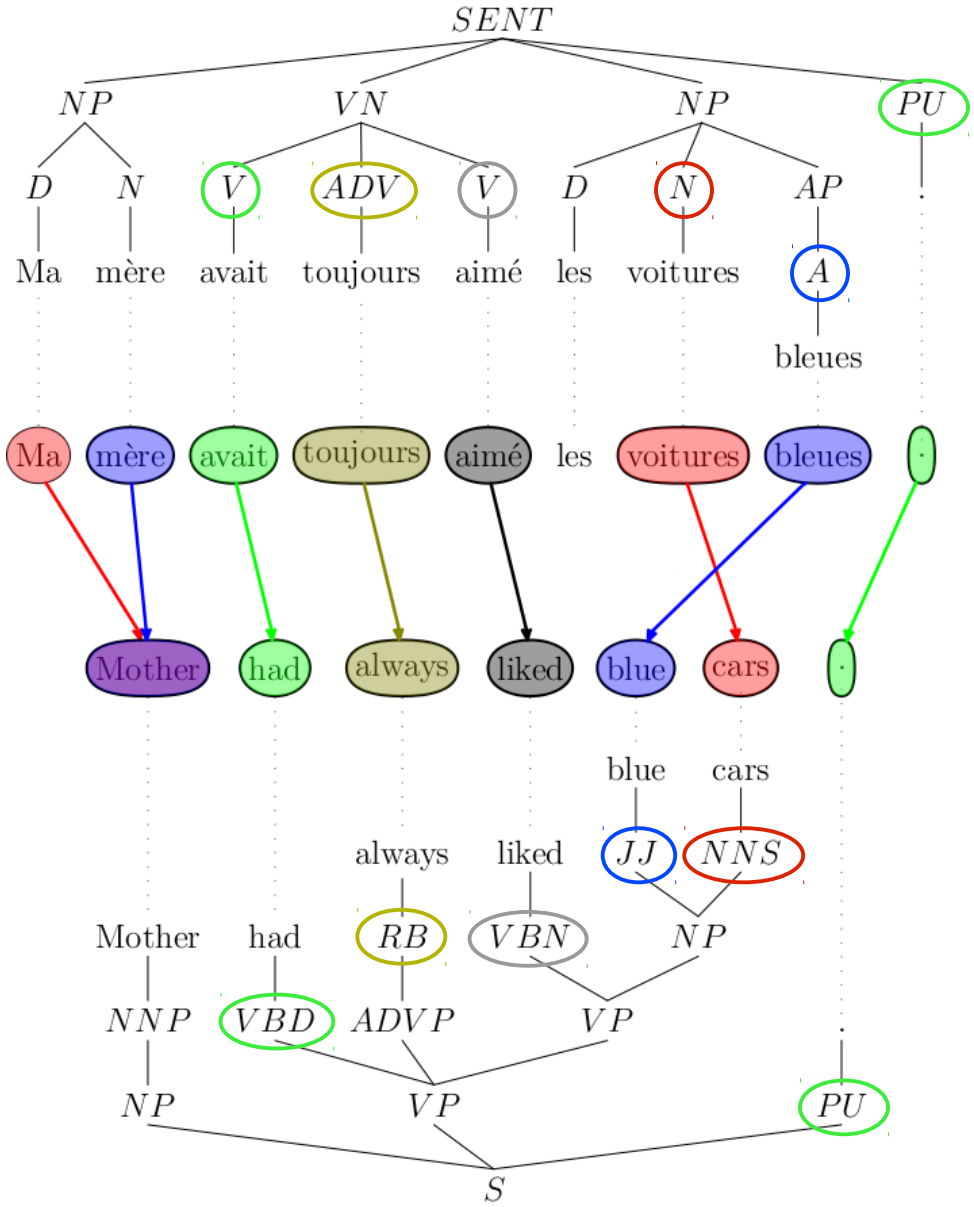
	BLEU	
Best Syntax-Only Baseline	24.1	+2.1
Best Overall Baseline	26.4	+1.2
Best WMT 2010 System	29.6	

- Syntax-only better for grammar analysis
- Expected improvements from improved grammar pruning and rule extraction

Computing Demands



Baseline Extraction Method



V::VBD → [avait]::[had]

ADV::RB → [toujours]::[always]

V::VBN → [aimé]::[liked]

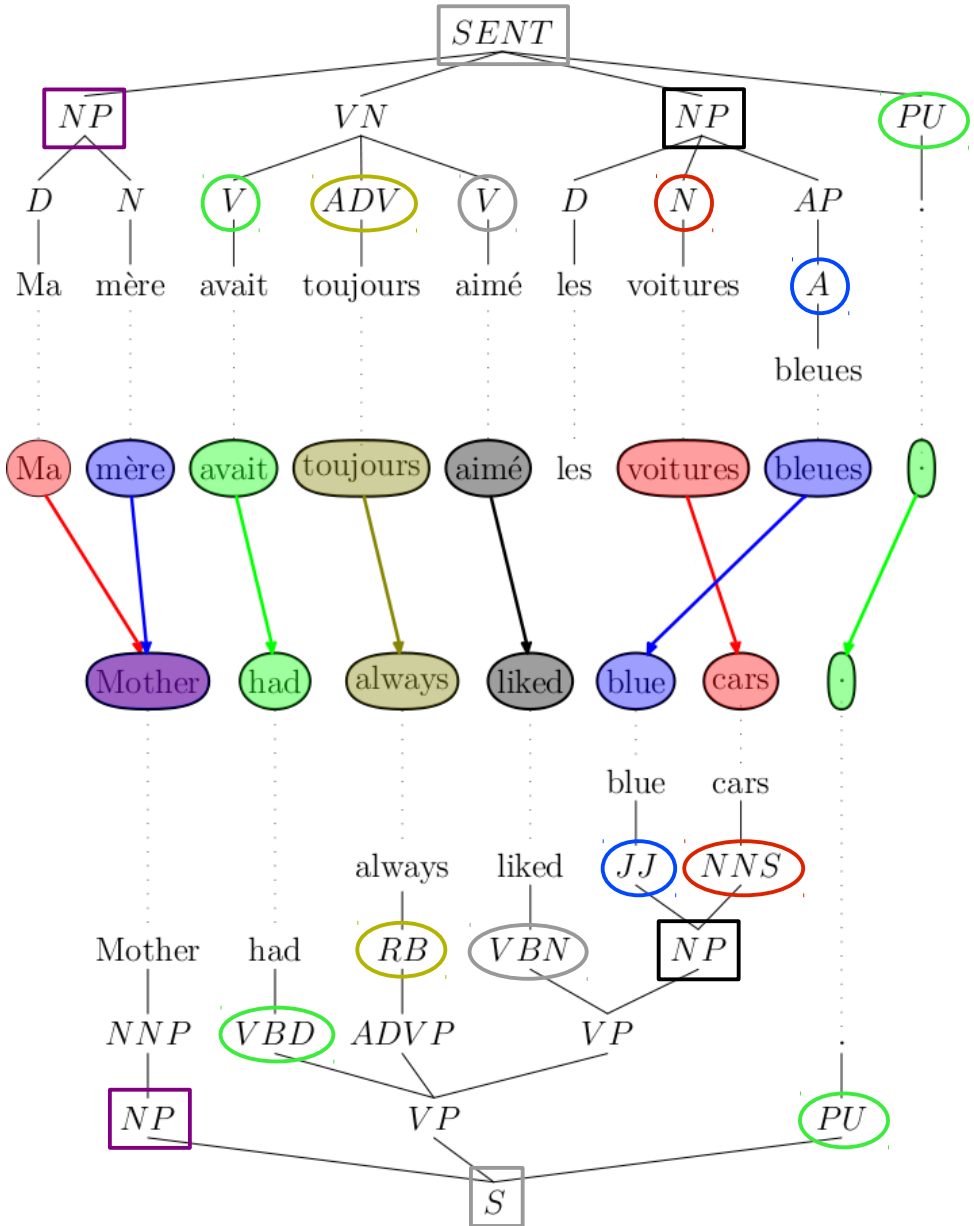
N::NNS → [voitures]::[cars]

A::JJ → [bleues]::[blue]

PU::PU → [.]::[.]

[Lavie et al. 2008]

Baseline Extraction Method



NP::NP → [Ma mère]::
[Mother]

NP::NP → [les voitures bleues]::
[blue cars]

NP::NP → [les N^1 A^2]::
[JJ^2 NNS^1]

SENT::S →
[NP^1 V^2 ADV^3 V^4 NP^5 PU^6]::
[NP^1 VBD^2 RB^3 VBN^4 NP^5 PU^6]

[Lavie et al. 2008]