

Inductive Detection of Language Features via Clustering Minimal Pairs

Toward Feature-Rich Grammars in Machine Translation

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA



ACL SSST-2 - June 20, 2008



Outline

- Overview of Feature Detection
- Example Application: Feature-Rich Grammars
- The Process of Feature Detection
- Results
- Conclusions



Feature Detection

((num sg)...)

The dog eats

El perro come

((num dl)...)

The dogs eat

Los perros comen

((num pl)...)

The dogs eat

Los perros comen



Feature Detection

((num sg)...)

The dog eats

El perro come

((num dl)...)

The dogs eat

Los perros comen

((num pl)...)

The dogs eat

Los perros comen

context: There are two dogs



Feature Detection

((num sg)...)

The dog eats

El perro come

((num dl)...)

The dogs eat

Los perros comen

((num pl)...)

The dogs eat

Los perros comen



Feature Detection

((num sg)...)

The dog eats

El perro come

((num dl)...)

The dogs eat

Los perros comen

((num pl)...)

The dogs eat

Los perros comen

Does this language distinguish singular,
dual, or plural agents?

If so, how?



Feature Detection

((num sg)...)

The dog eats

El perro come

((num dl)...)

The dogs eat

Los perros comen

((num pl)...)

The dogs eat

Los perros comen

Feature
Detection

Feature	Candidate Lexical Items
(num sg)	come, el, perro
(num dl, pl)	comen, los, perros

Does this language distinguish singular,
dual, or plural agents?

If so, how?



Outline

- ✓ Overview of Feature Detection
 - Example Application: Feature-Rich Grammars
 - The Process of Feature Detection
 - Results
 - Conclusions



Problem: Long-Distance Interactions

English → Urdu

“A student named Irshad who was throwing flour in the water for the fish . . .”

ek	talb	alm	arshad	jo	mchhlyoN	ke liye
a.SG	student	named	Irshad	who	fish	for
← 12 words →						
pani	maiN	aata	phink	raha	tha	...
water	in	flour	throw	PROG.SG.M	be.PAST.SG.M	



Example Application: Feature-Rich Grammars

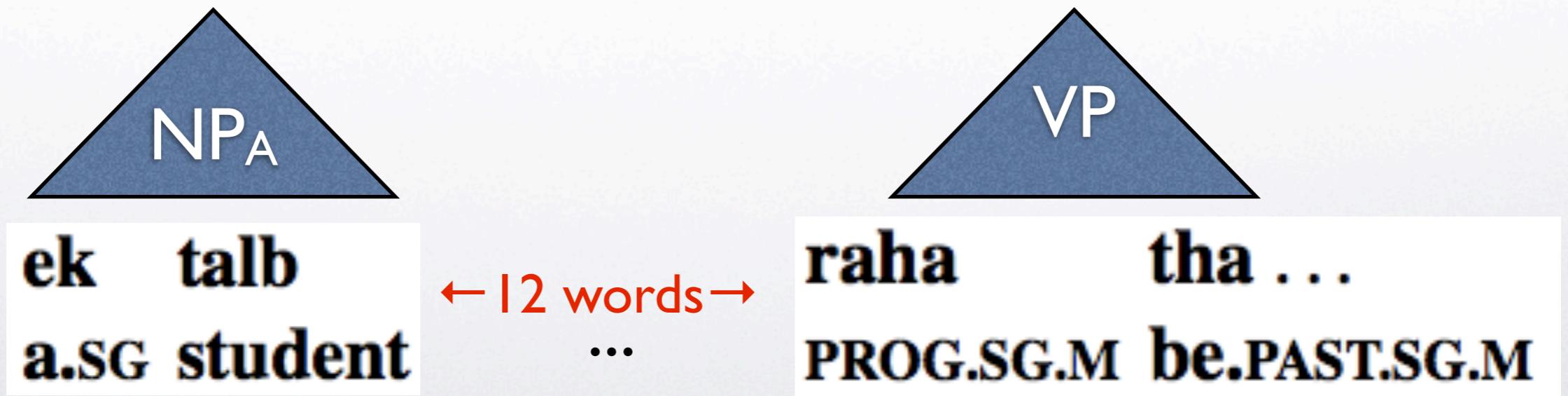
ek talb
a.SG student

← 12 words →
...

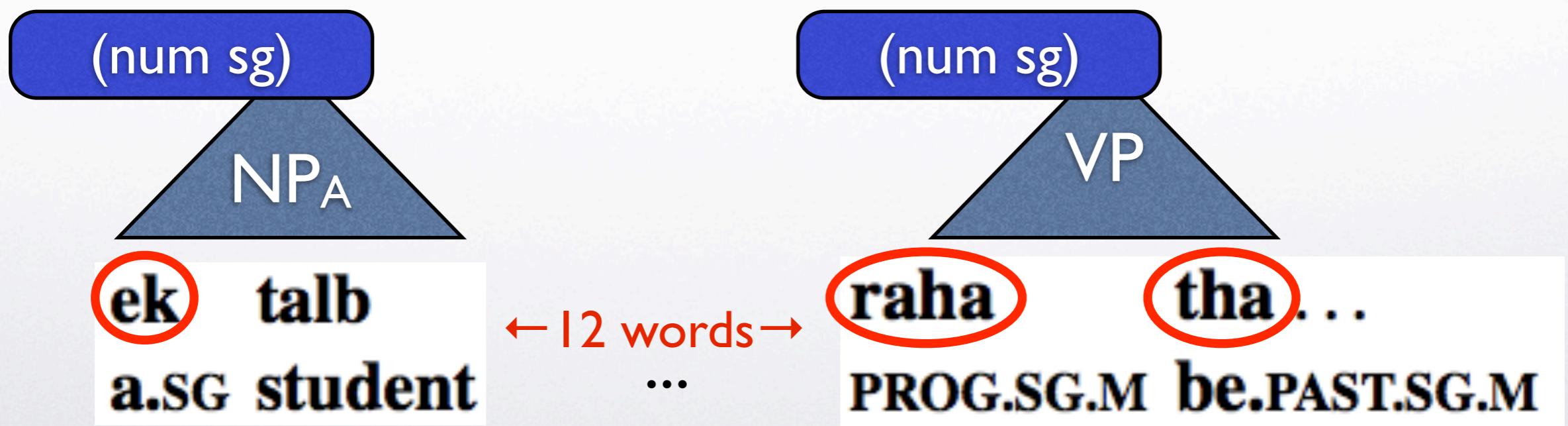
raha tha ...
PROG.SG.M be.PAST.SG.M



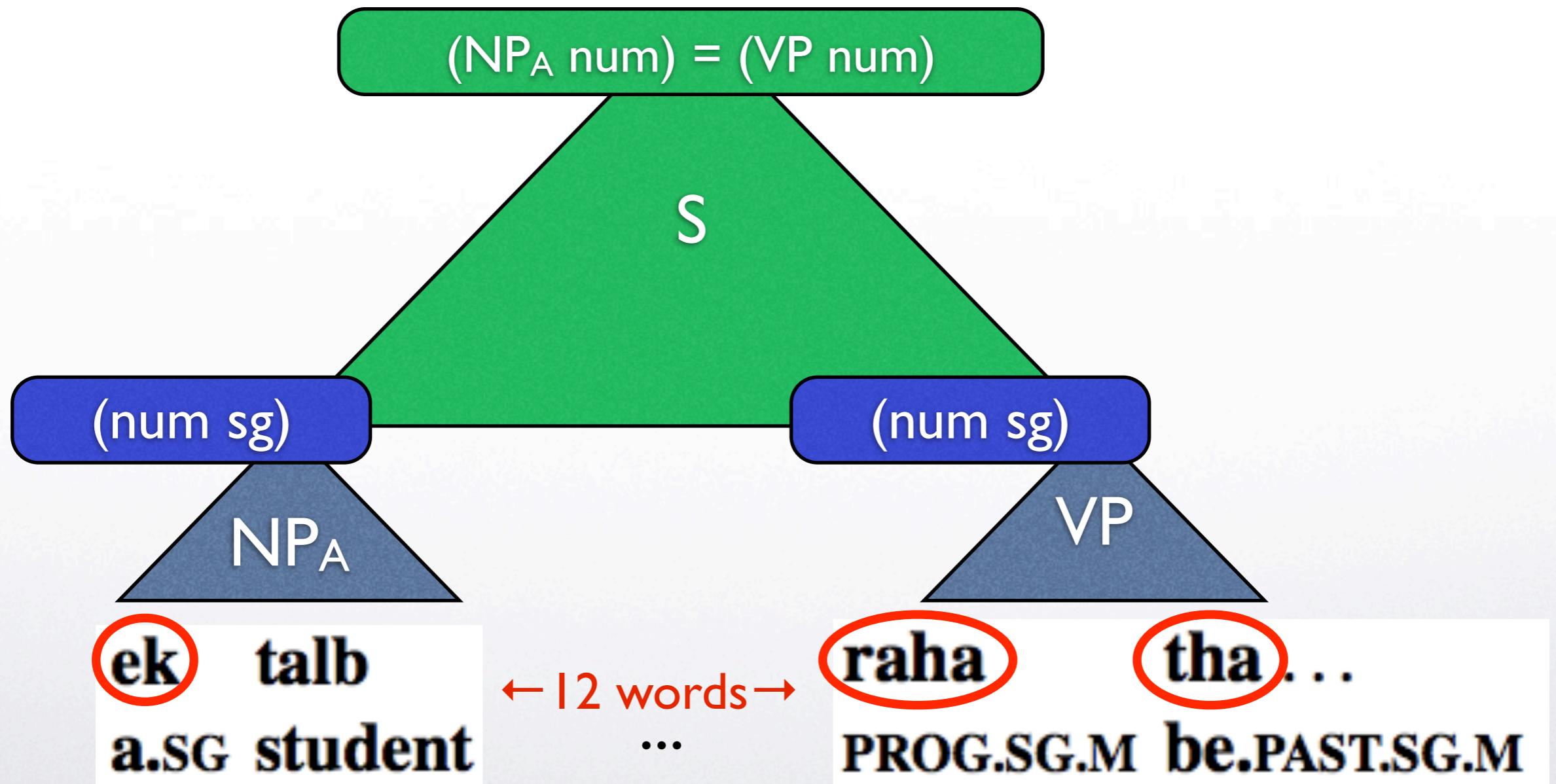
Example Application: Feature-Rich Grammars



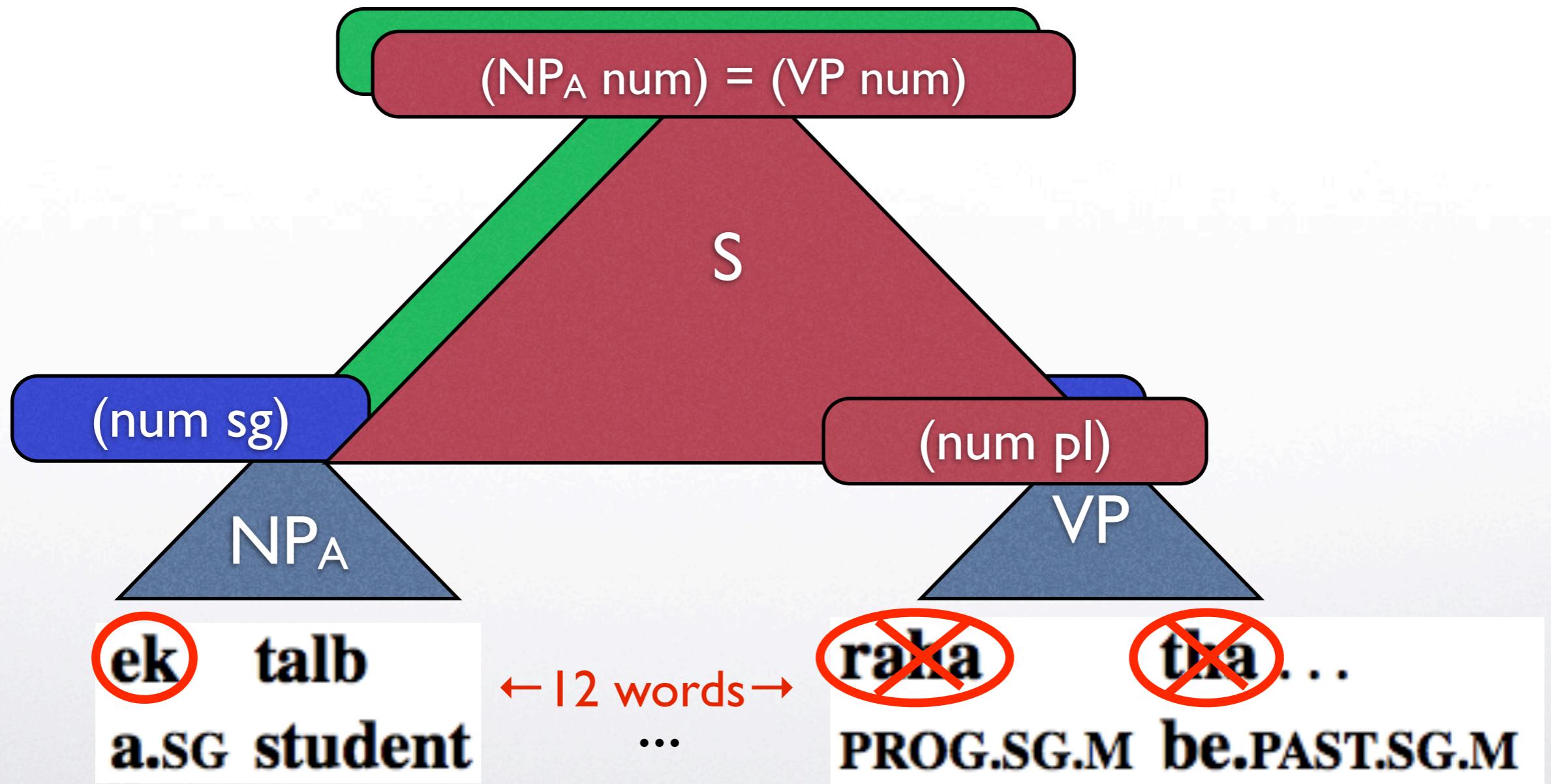
Example Application: Feature-Rich Grammars



Example Application: Feature-Rich Grammars



Example Application: Feature-Rich Grammars

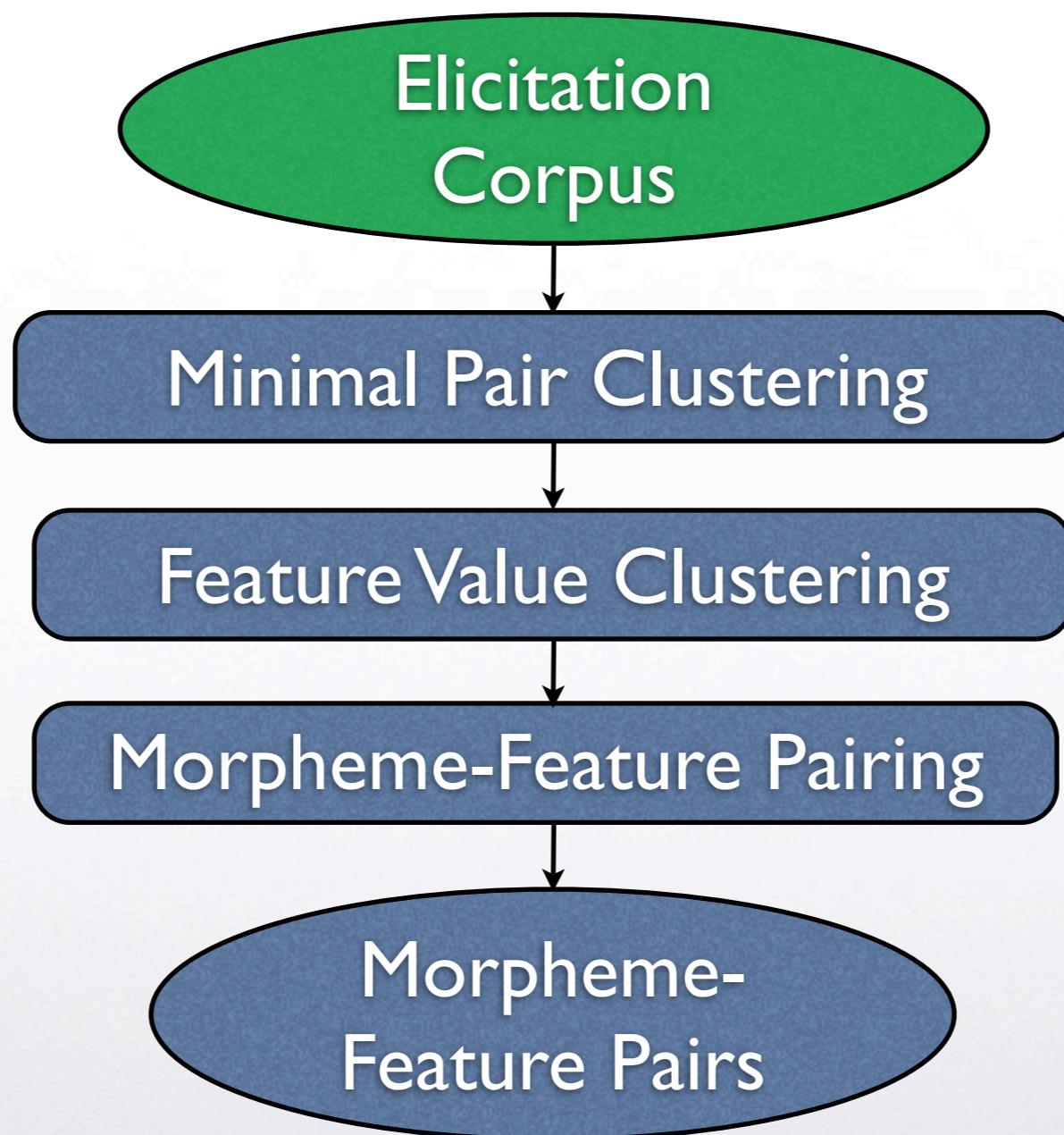


Outline

- ✓ Overview of Feature Detection
- ✓ Example Application: Feature-Rich Grammars
- The Process of Feature Detection
- Results
- Conclusions



Inductive Feature Detection



Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies .



Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies .

AVENUE Elicitation Tool V1.56 English-Urdu

File Edit Tools Help

<< < 37 > >> Add Alternate Translation Text - Text +

Context: Translate this sentence as if the incident it refers to happened weeks ago.

Eliciting: Did Amna not give Danish books?

Add Phrase

Elicited: کیا آمنہ نے دانش کو کتابیں نہیں دیں؟

Alignment: ((1,1),(2,2),(3,7),(4,8),(5,4),(6,6))

Comment:

The screenshot shows the AVENUE Elicitation Tool V1.56 English-Urdu interface. The window title is "AVENUE Elicitation Tool V1.56 English-Urdu". The menu bar includes "File", "Edit", "Tools", and "Help". Below the menu is a toolbar with buttons for navigating through entries: "<<" (left), "<" (previous), "37" (current page), ">" (next), ">>" (right), "Add Alternate Translation", "Text -" (decrease text size), and "Text +" (increase text size). The main area is divided into five sections: "Context", "Eliciting", "Elicited", "Alignment", and "Comment". The "Context" section contains the instruction "Translate this sentence as if the incident it refers to happened weeks ago." The "Eliciting" section displays the sentence "Did Amna not give Danish books?" with individual words highlighted in colored boxes (green, pink, orange, blue) and a red question mark. The "Elicited" section shows the translation "کیا آمنہ نے دانش کو کتابیں نہیں دیں؟" with individual words highlighted in colored boxes (orange, pink, blue, green). The "Alignment" section shows a list of word pairs: ((1,1),(2,2),(3,7),(4,8),(5,4),(6,6)). The "Comment" section is currently empty. There are scroll bars on the right side of each section panel.



Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies .
tgtsent: Maria hornea galletas .
aligned: ((1,1),(2,2),(3,3),(4,4))



Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.

srcsent: Maria bakes cookies .

tgtsent: Maria hornea galletas .

aligned: ((1,1),(2,2),(3,3),(4,4))

fstruct: [f1]([f2](actor ((gender f)(anim human)(num sg)))
[f3](undergoer ((person 3) (num dl))) (tense pres))



Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.

srcsent: Maria bakes cookies .

tgtsent: Maria hornea galletas .

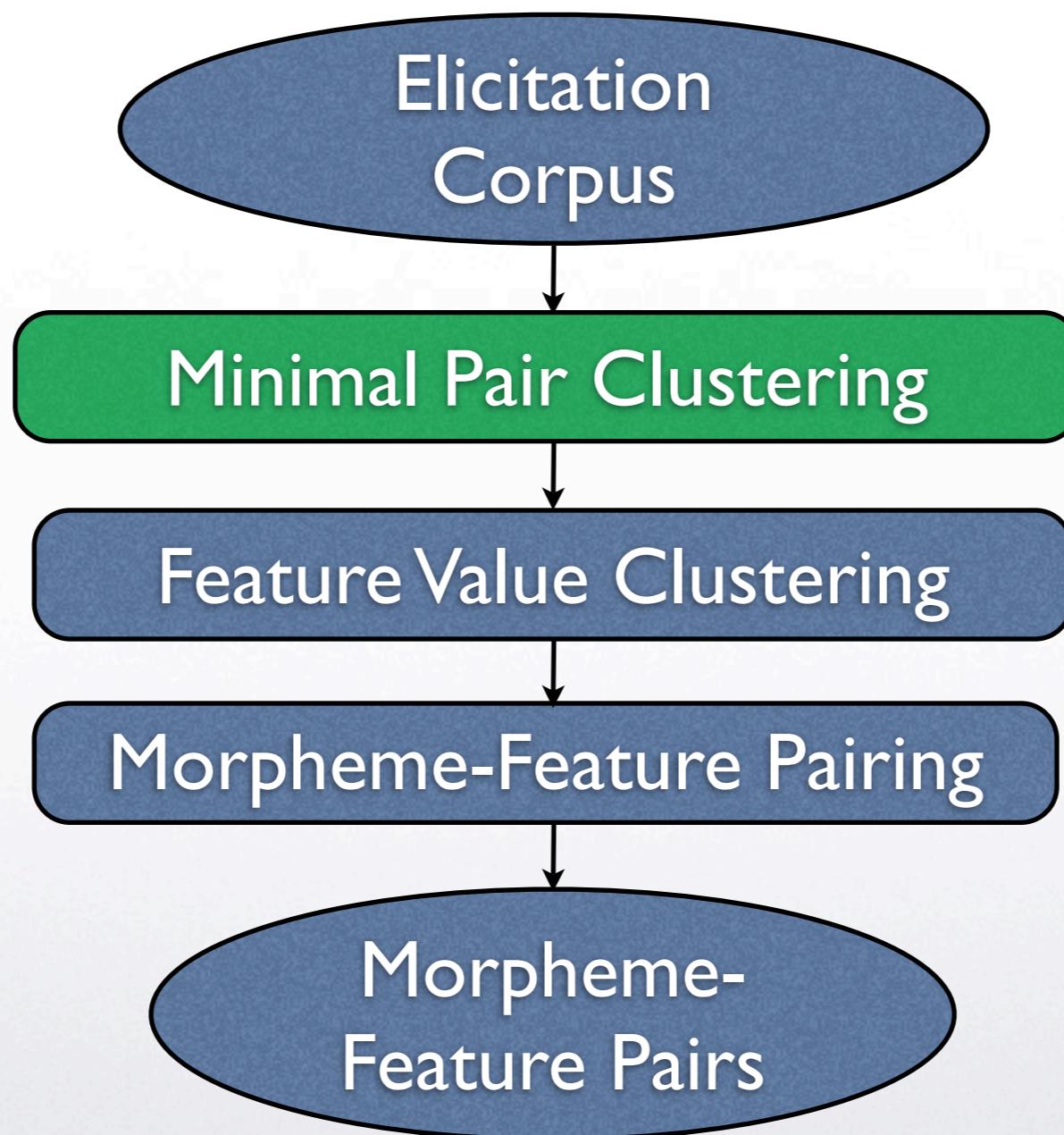
aligned: ((1,1),(2,2),(3,3),(4,4))

fstruct: [f1]([f2](actor ((gender f)(anim human)(num sg)))
[f3](undergoer ((person 3) (num dl))) (tense pres))

Distributed in this year's NIST Urdu MT
Evaluation via the LDC's LCTL Language Pack



Inductive Feature Detection



Minimal Pair Clustering

Wildcard Feature Structure:

((actor (num *) (gender n))

(undergoer (num sg) (gender n))

(tense pres))

sg

pl

The dog sees the cat

El perro ve el gato

((actor (num sg) (gender n))

(undergoer (num sg) (gender n))

(tense pres))

The dogs see the cat

Los perros ven el gato

((actor (num pl) (gender n))

(undergoer (num sg) (gender n))

(tense pres))



Minimal Pair Clustering

Wildcard Feature Structure:

((actor (num *) (gender n))

(undergoer (num sg) (gender n))

(tense pres))



The dog sees the cat

El perro ve el gato

((actor (num sg) (gender n))

(undergoer (num sg) (gender n))

(tense pres))

The dogs see the cat

Los perros ven el gato

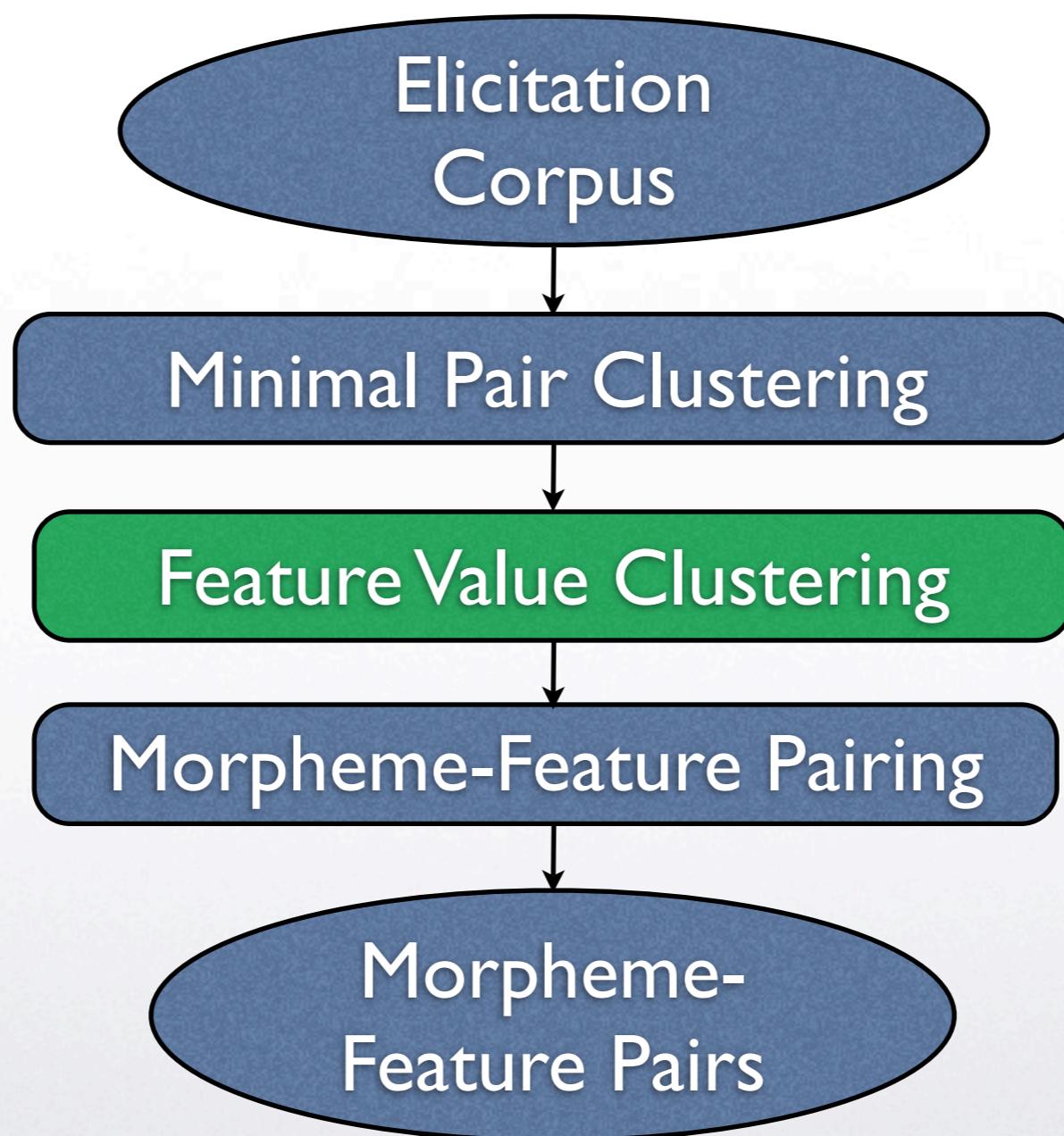
((actor (num pl) (gender n))

(undergoer (num sg) (gender n))

(tense pres))



Inductive Feature Detection



Feature Value Clustering

s I: The dog sees the cat

El perro ve el gato

((actor (num sg)...)))

sg

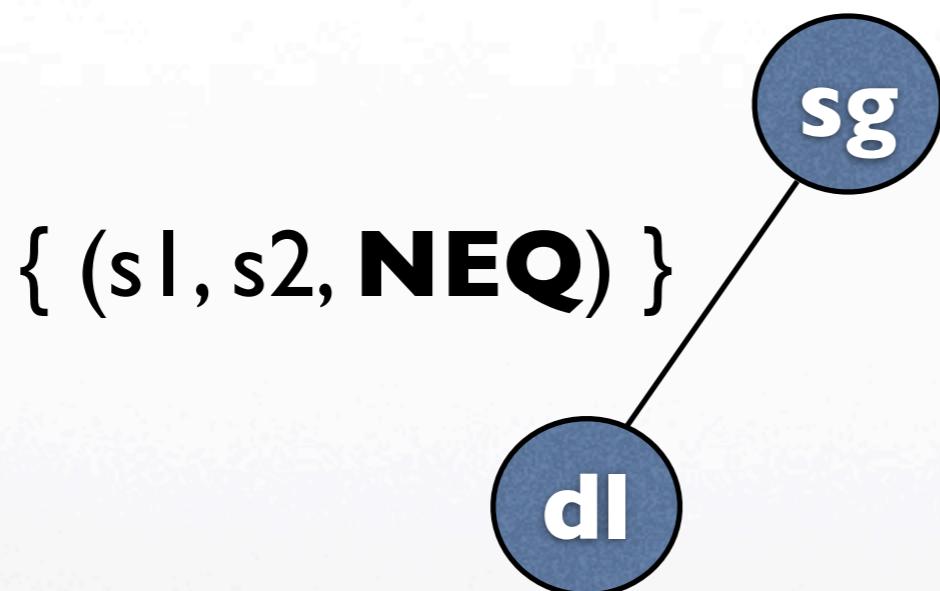


Feature Value Clustering

s1: The dog sees the cat

El perro ve el gato

((actor (num sg)...)))



s2:

The dogs see the cat

Los perros ven el gato

((actor (num dl)...)))

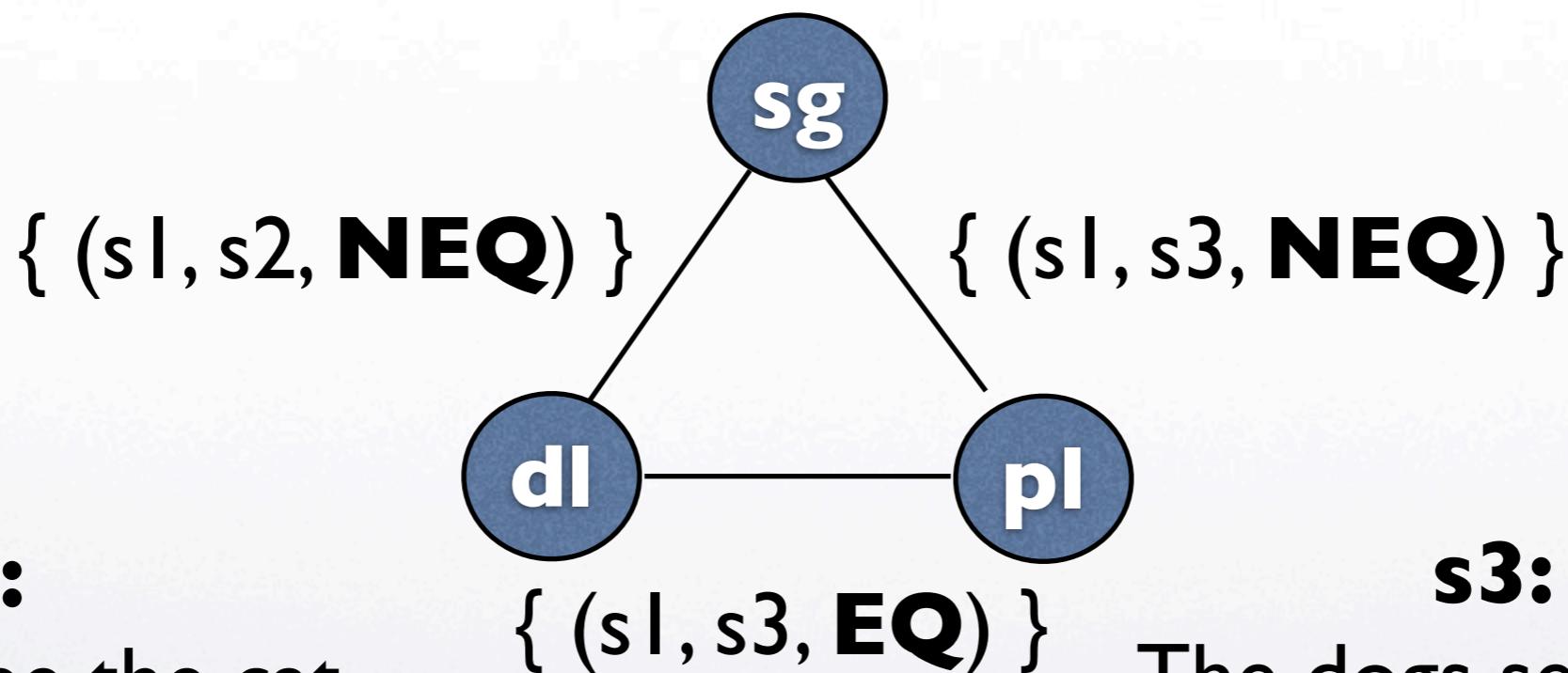


Feature Value Clustering

s1: The dog sees the cat

El perro ve el gato

((actor (num sg)...)))



s2:

The dogs see the cat

Los perros ven el gato

((actor (num dl)...)))

s3:

The dogs see the cat

Los perros ven el gato

((actor (num pl)...)))

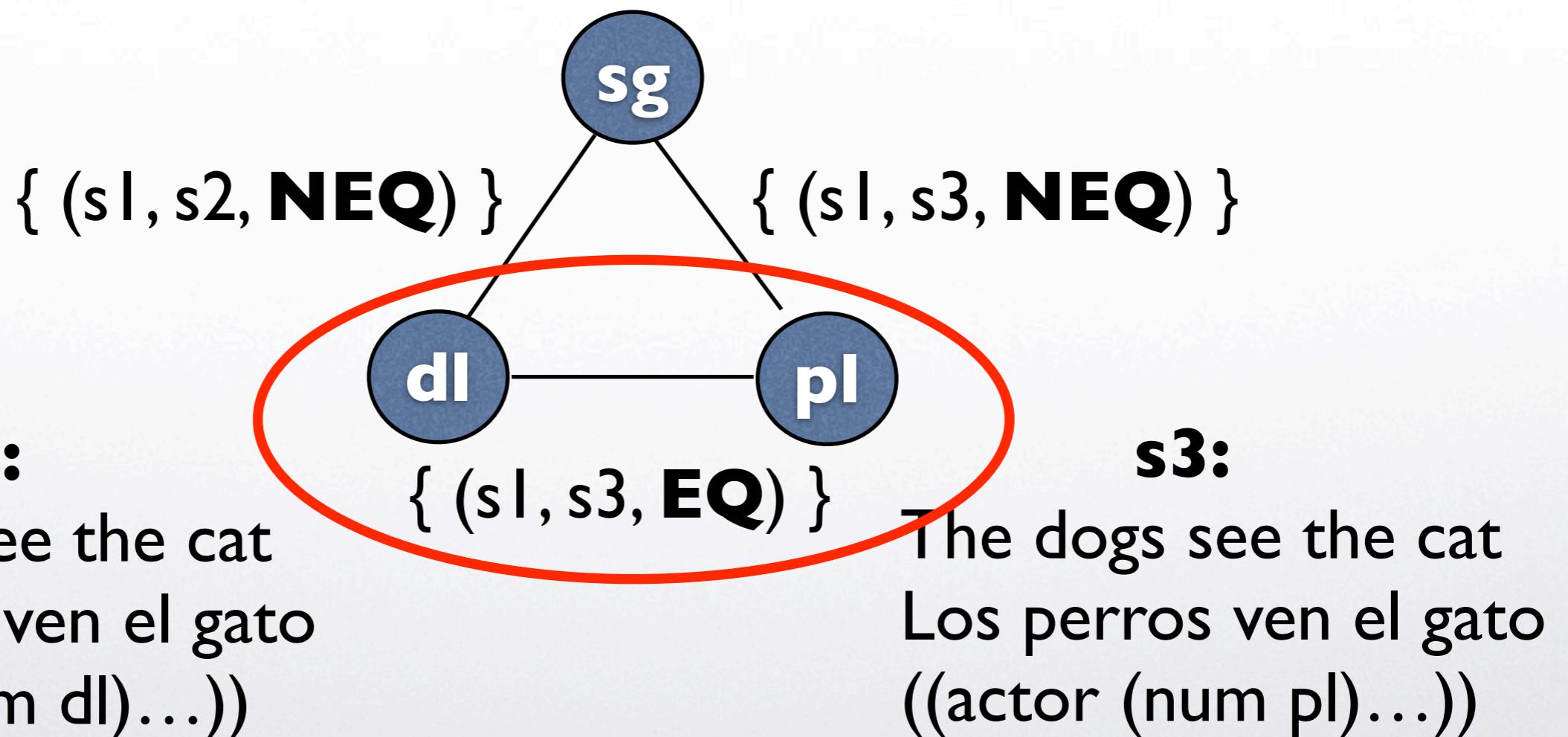


Feature Value Clustering

s1: The dog sees the cat

El perro ve el gato

((actor (num sg)...)))



Feature Value Clustering

s1: The dog sees the cat

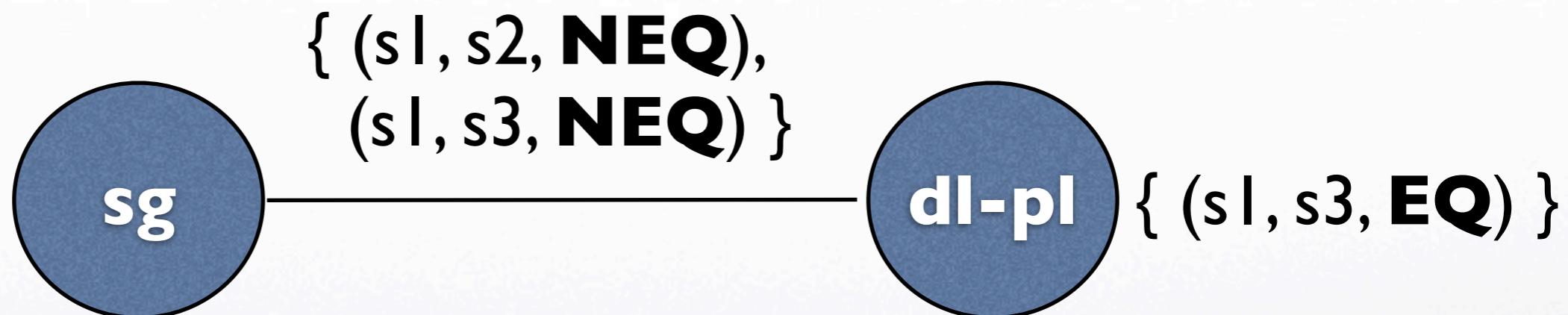
El perro ve el gato

((actor (num sg)...))

s2: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))



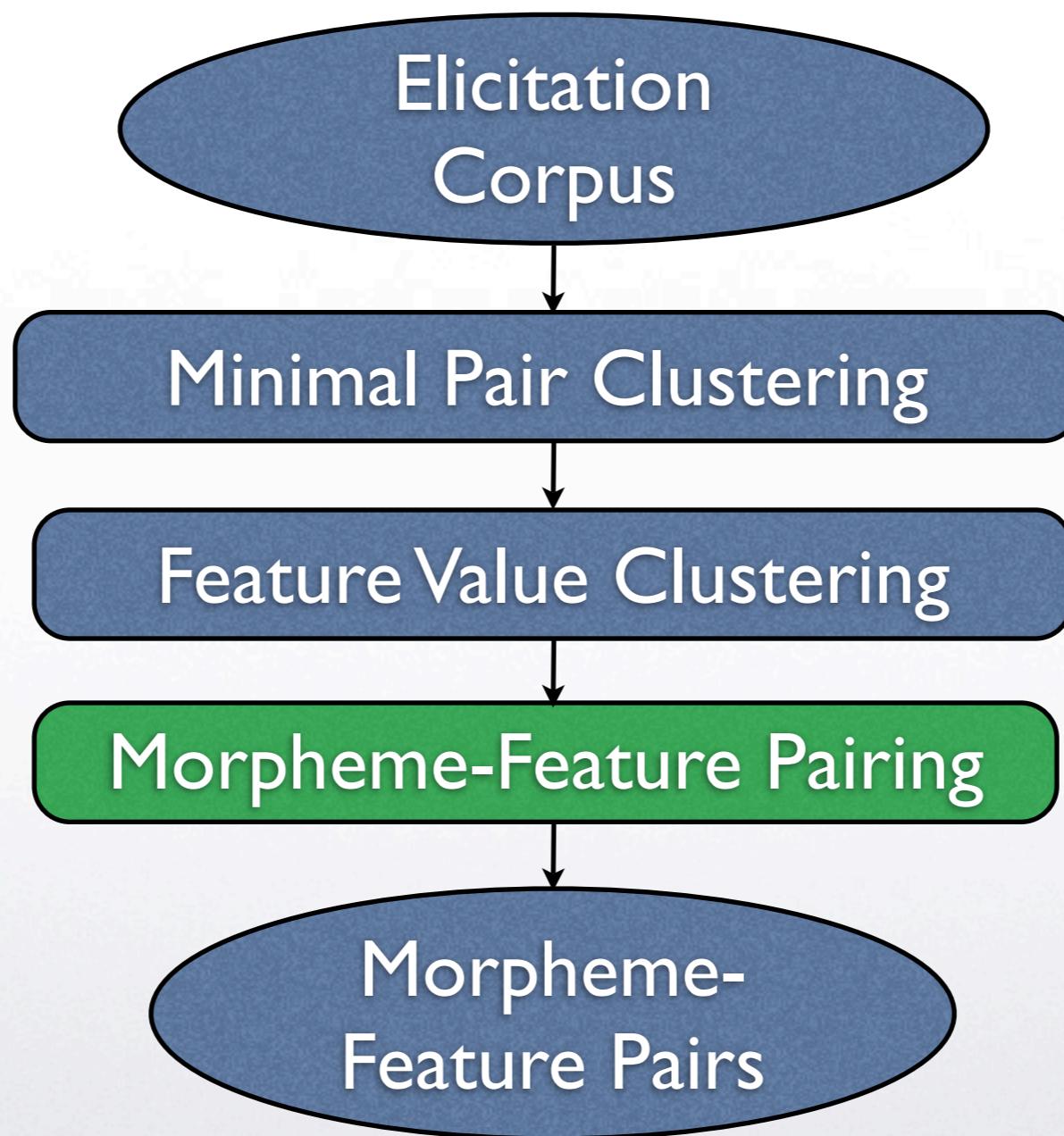
s3: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))



Inductive Feature Detection



Morpheme-Feature Pairing

s1: The dog sees the cat

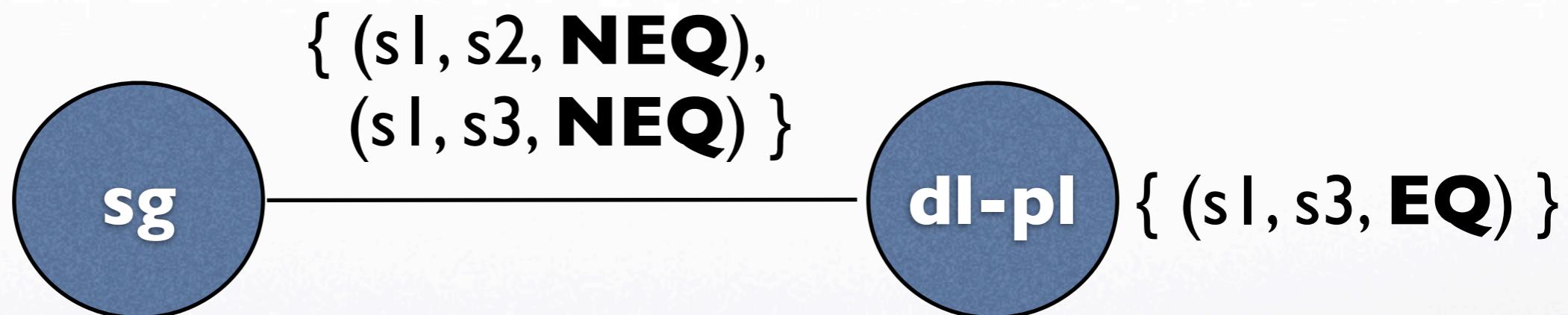
El perro ve el gato

((actor (num sg)...))

s2: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))



s3: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))



Morpheme-Feature Pairing

s1: The dog sees the cat

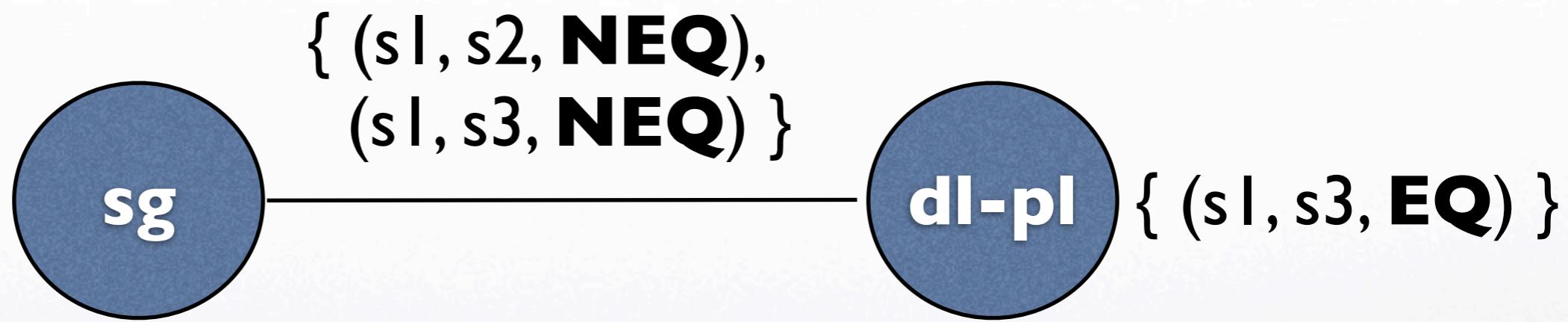
El perro ve el gato

((actor (num sg)...))

s2: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))



{ el, perro, ve }

{ los, perros, ven }

s3: The dogs see the cat

Los perros ven el gato

((actor (num dl)...))

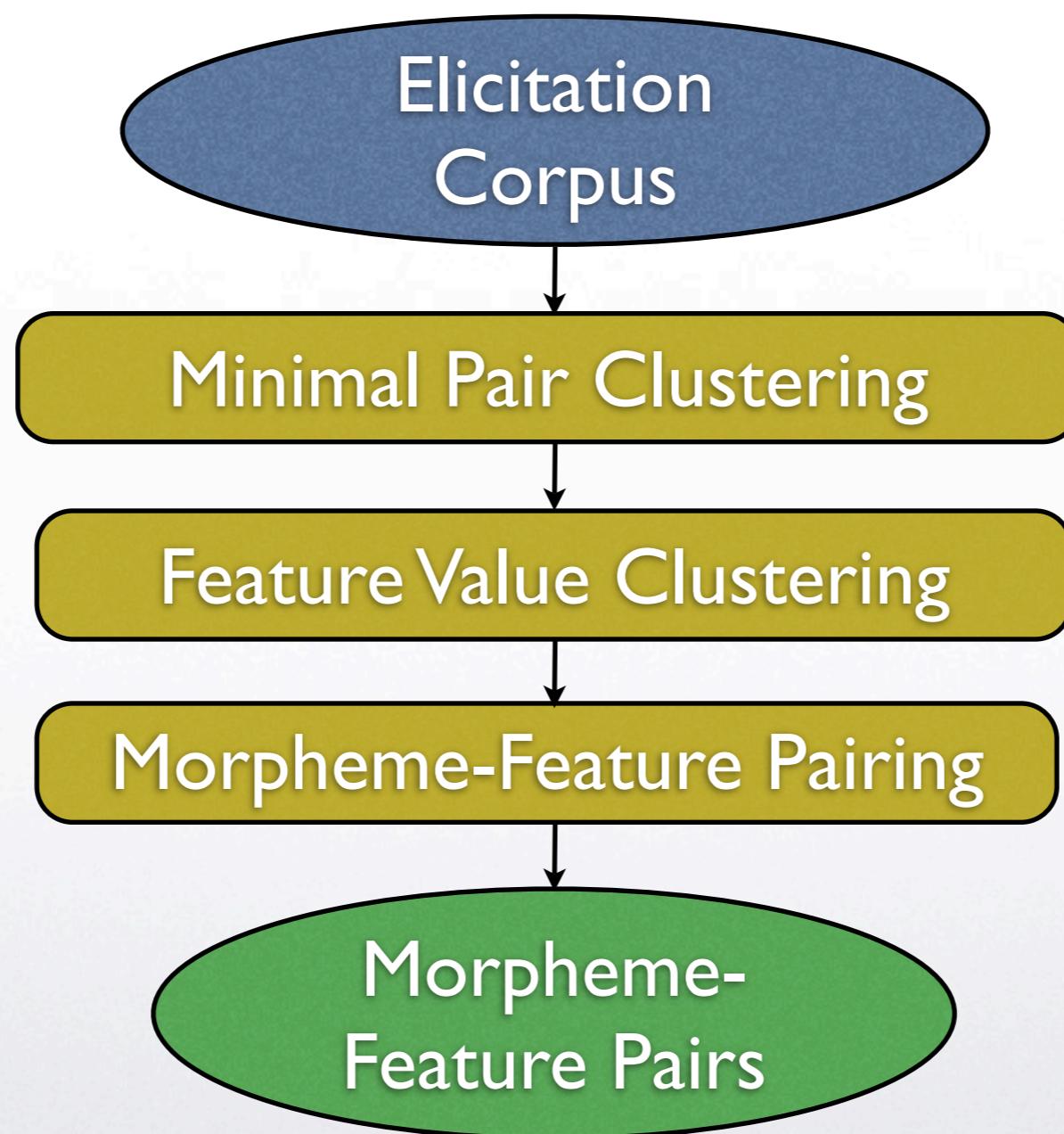


Outline

- ✓ Overview of Feature Detection
- ✓ Example Application: Feature-Rich Grammars
- ✓ The Process of Feature Detection
 - Results
 - Conclusions



Evaluation



Experiment

- Analyzed LCTL Urdu-English Elicitation Corpus (~3000 sentences)
- Evaluated by Urdu native speaker knowledgeable in linguistics

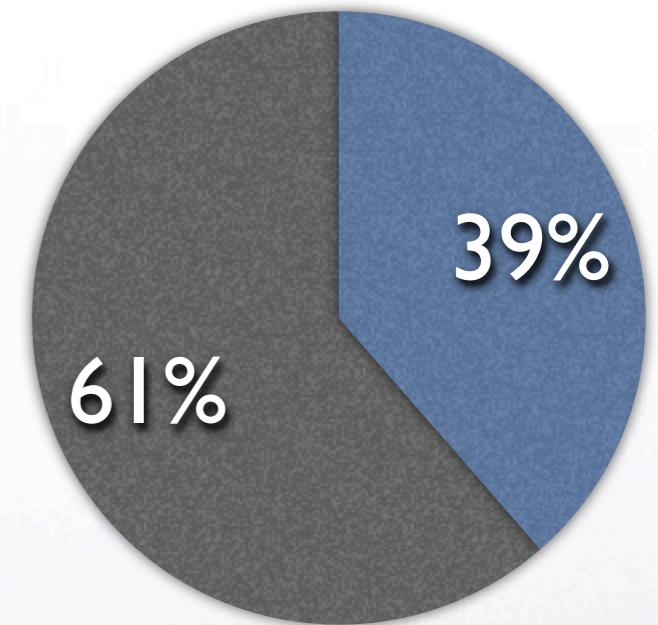
Judgement	Morpheme-Feature Pairings	Example Output
Correct	68	“hai” → (tense pres)
Ambiguous	29	“raha” → (tense pres)
Incorrect	109	“nahin” → (tense pres)
TOTAL	206	



Experiment

- Found **68 Correct Morpheme-Feature Pairs**
 - = 53 Word Types
 - In an Urdu corpus of **17M tokens** and 200k types, the 53 types selected by feature detection cover:
 - 0.02% of Types (Vocabulary)
 - **38.6% of Tokens**

Tokens in
Blind Test Set



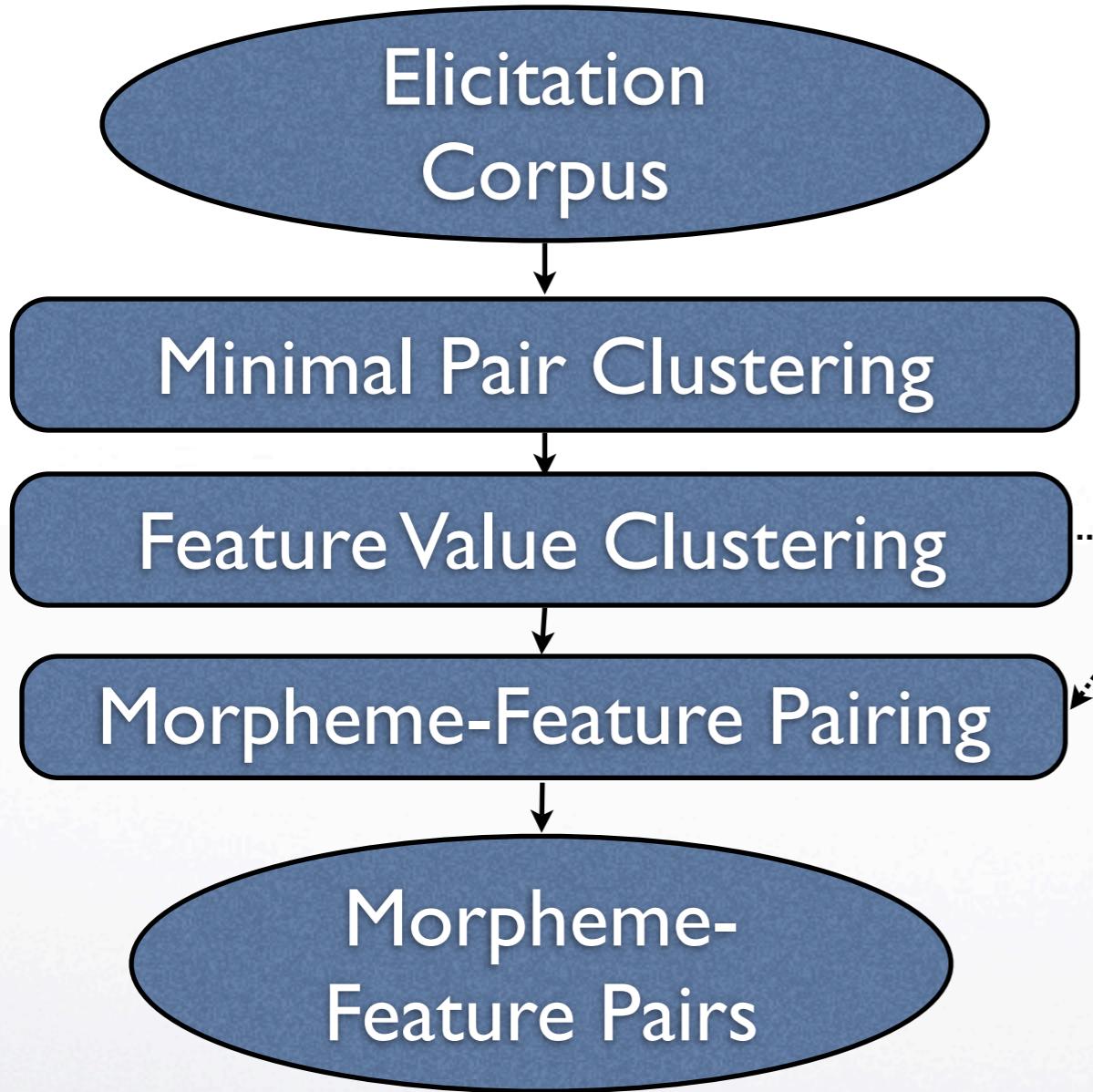
● Annotated
● Remaining



Outline

- ✓ Overview of Feature Detection
- ✓ Example Application: Feature-Rich Grammars
- ✓ The Process of Feature Detection
- ✓ Results
- Conclusions

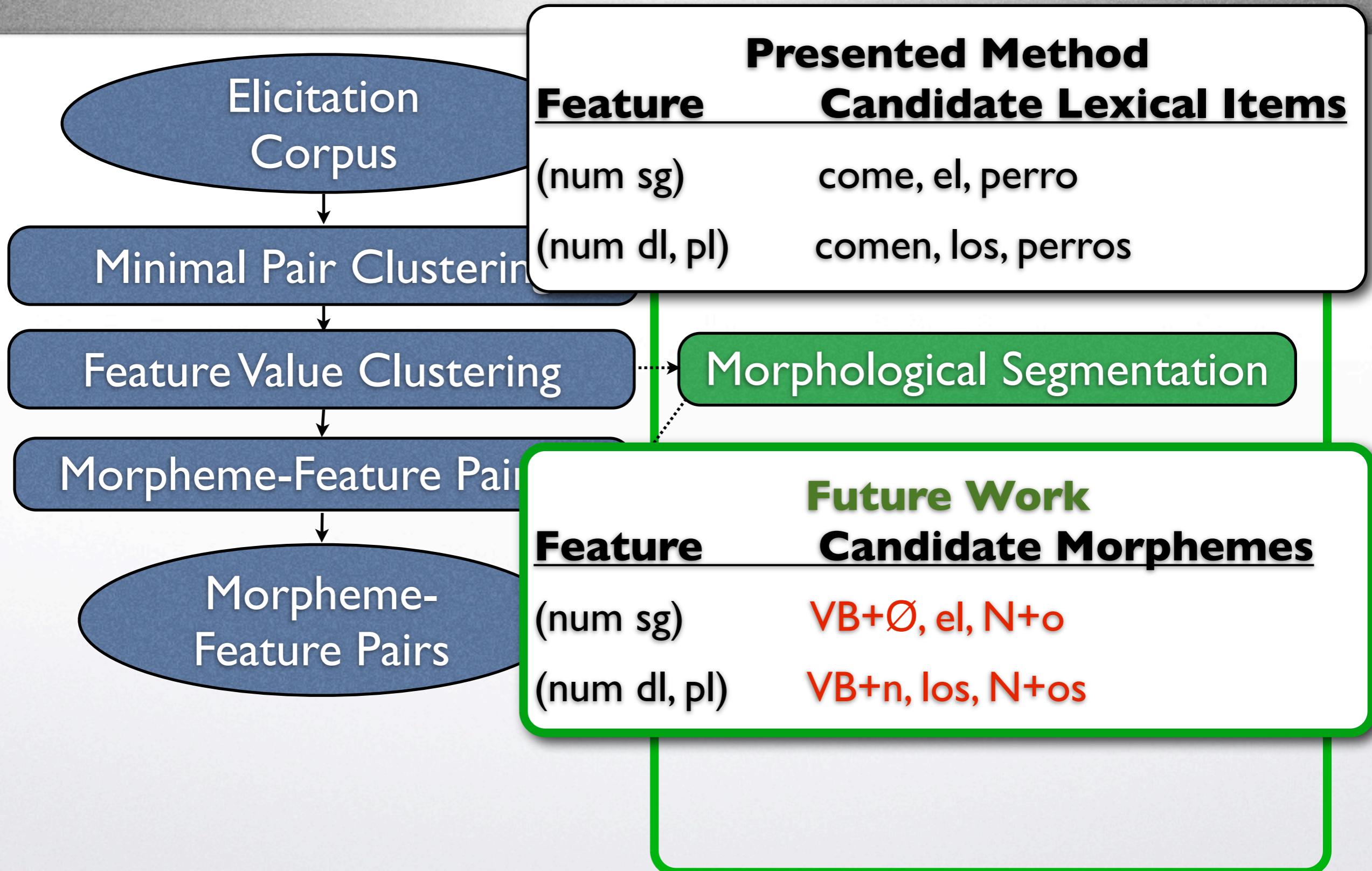


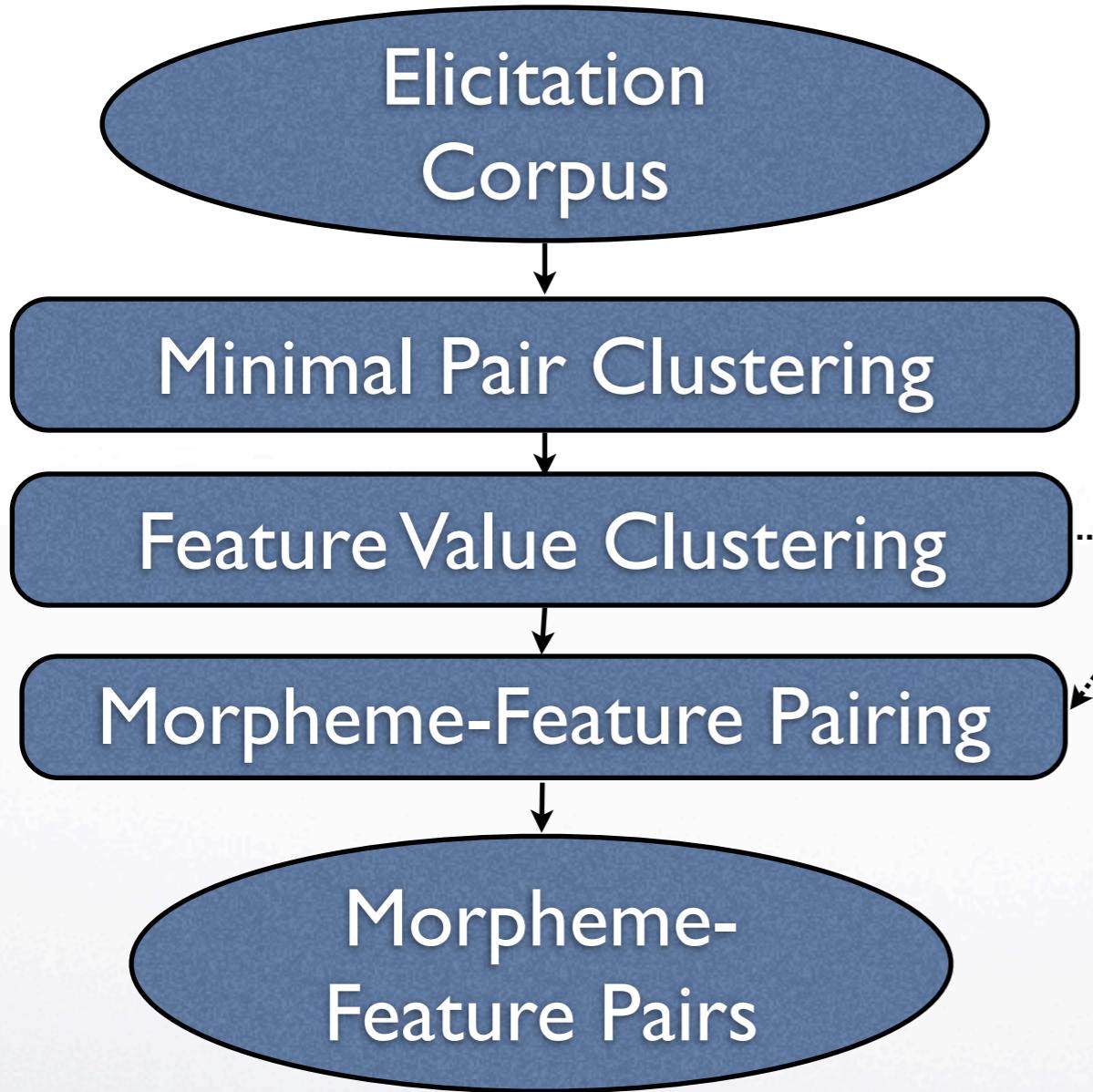


Future Work

Morphological Segmentation



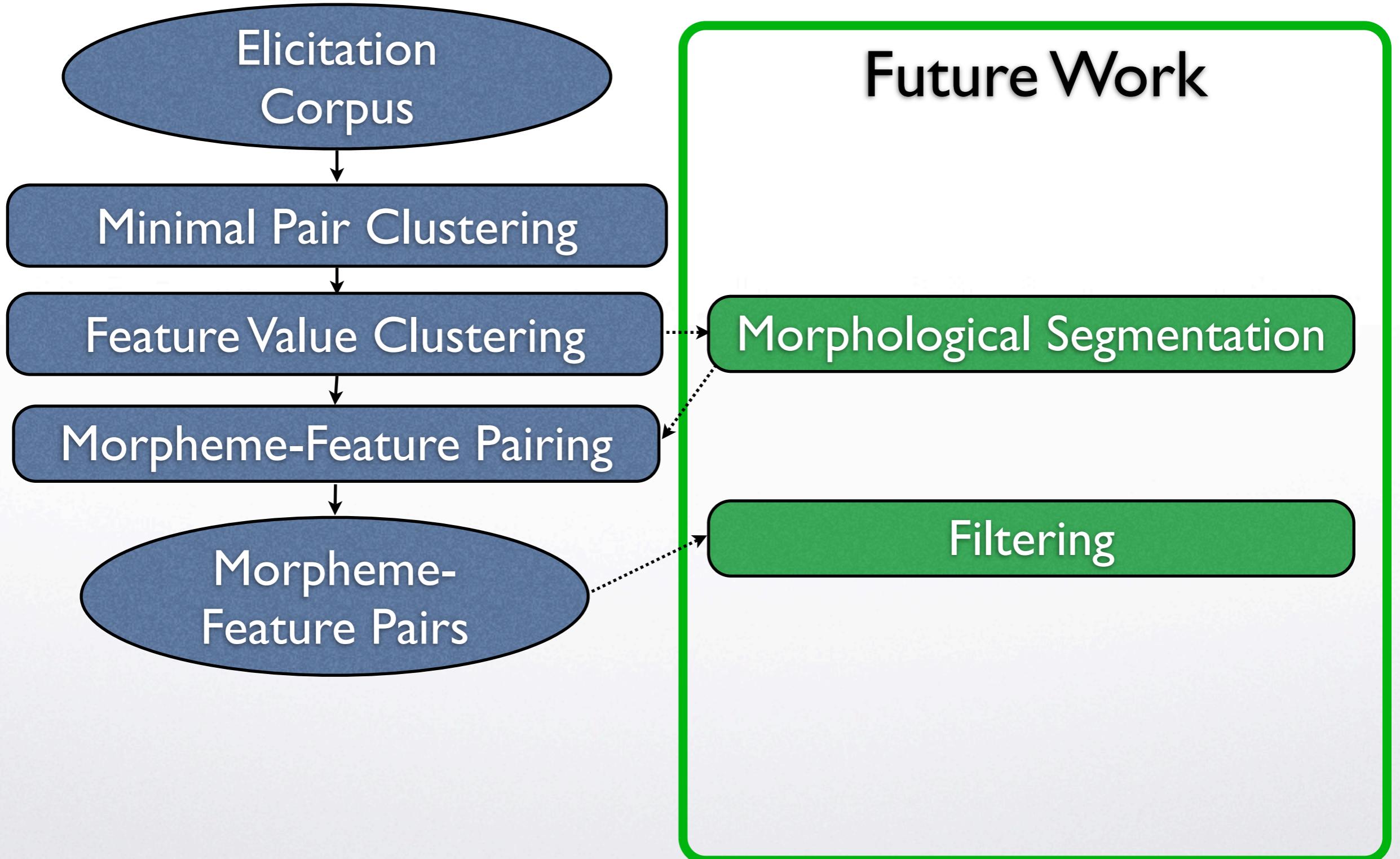


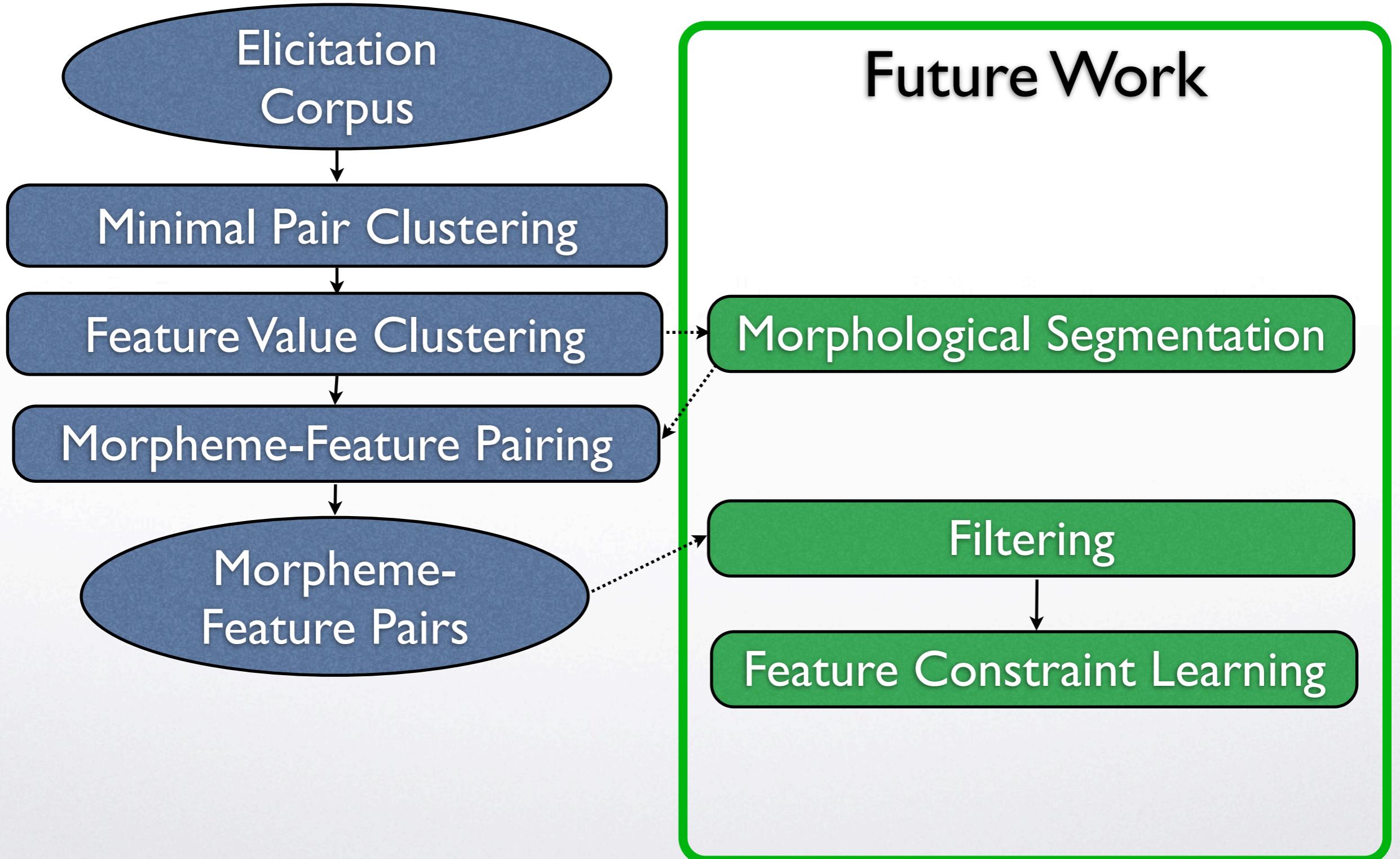


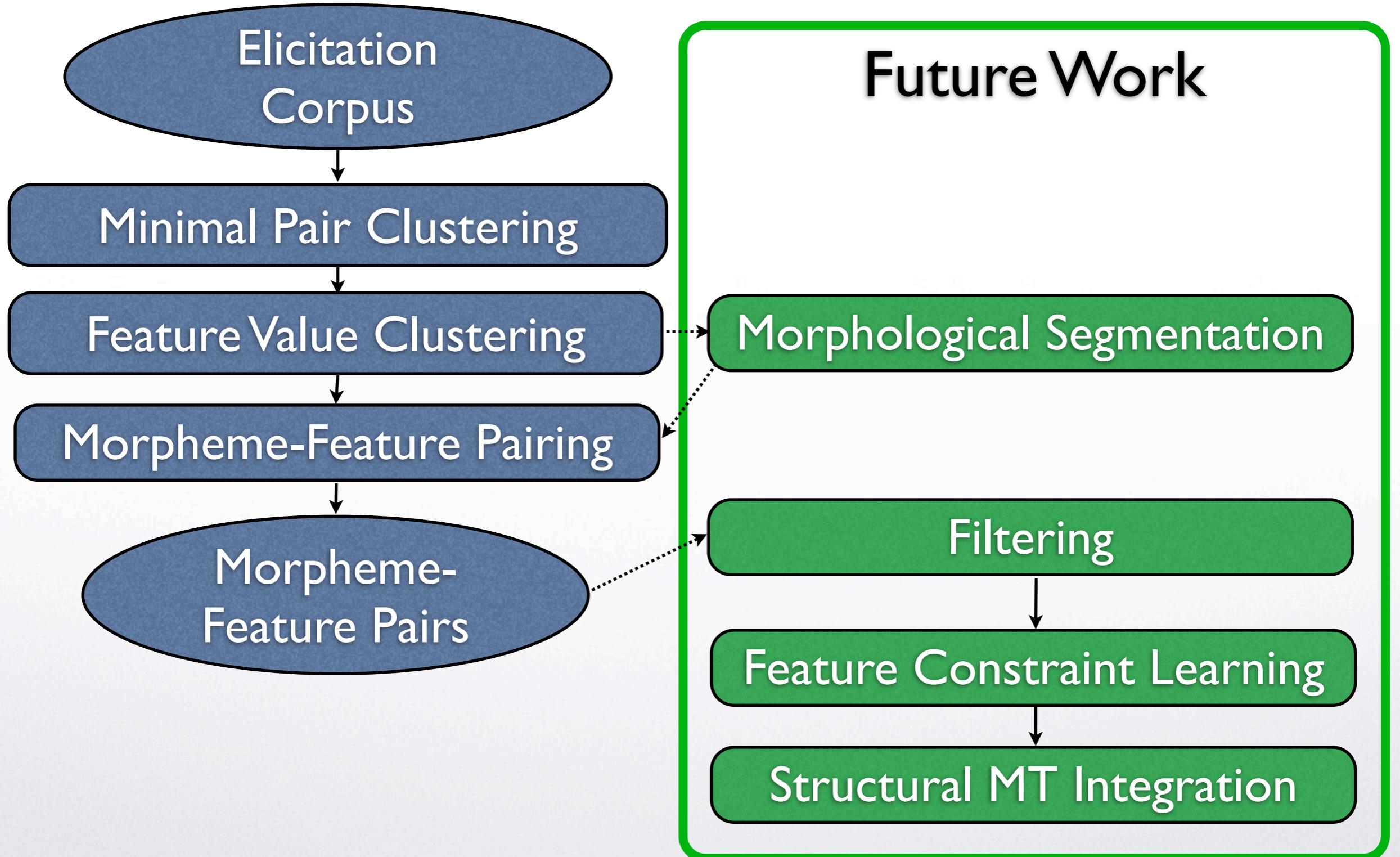
Future Work

Morphological Segmentation









Other Applications

- Factored MT
- Data Selection via active learning for synchronous grammar induction
- Aid for linguistics field work



Conclusion

- We now have
 - Feature-annotations for lexical items that convey grammatical meanings
 - Significant coverage
- Structural MT systems stand to benefit by incorporating this morphosyntactic information



Inductive Detection of Language Features via Clustering Minimal Pairs:

Toward Feature-Rich Grammars in Machine Translation

Questions?

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA

