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?

lti

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→

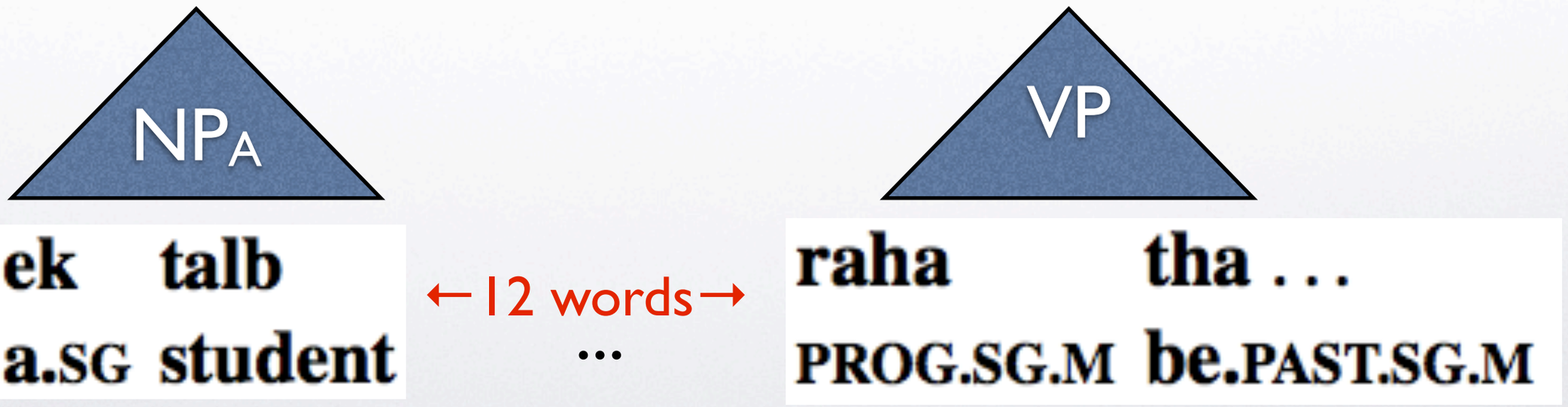
pani	maiN	aata	phink	**raha**	**tha** . . .	
water	in	flour	throw	**PROG.SG.M**	**be.PAST.SG.M**	

lti

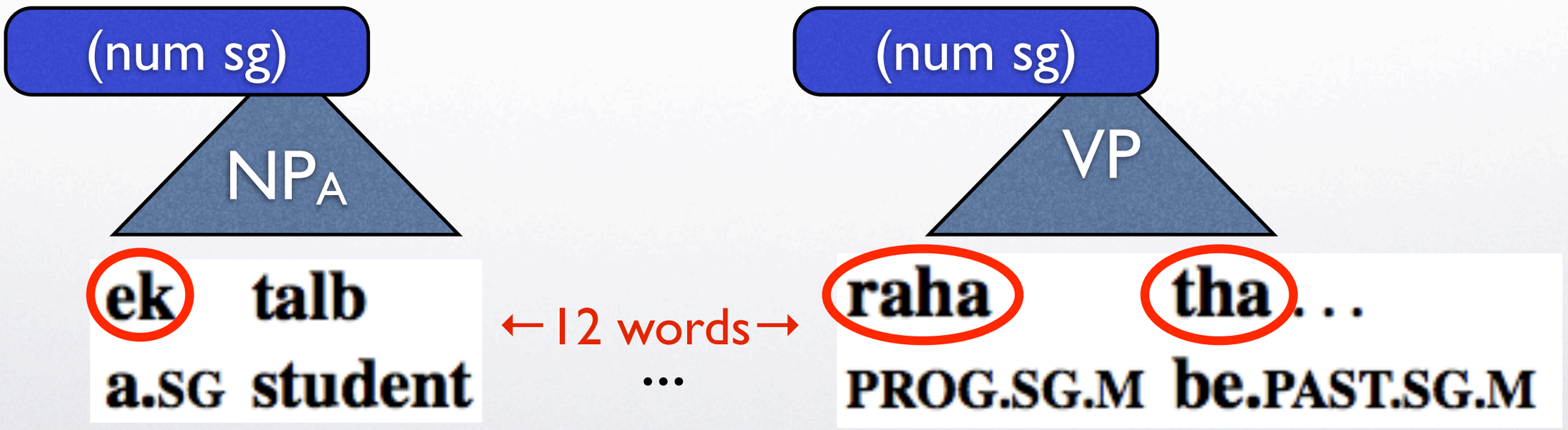
Example Application: Feature-Rich Grammars

ek	talb	← 12 words →		raha	tha . . .
a.SG	student	...		PROG.SG.M	be.PAST.SG.M

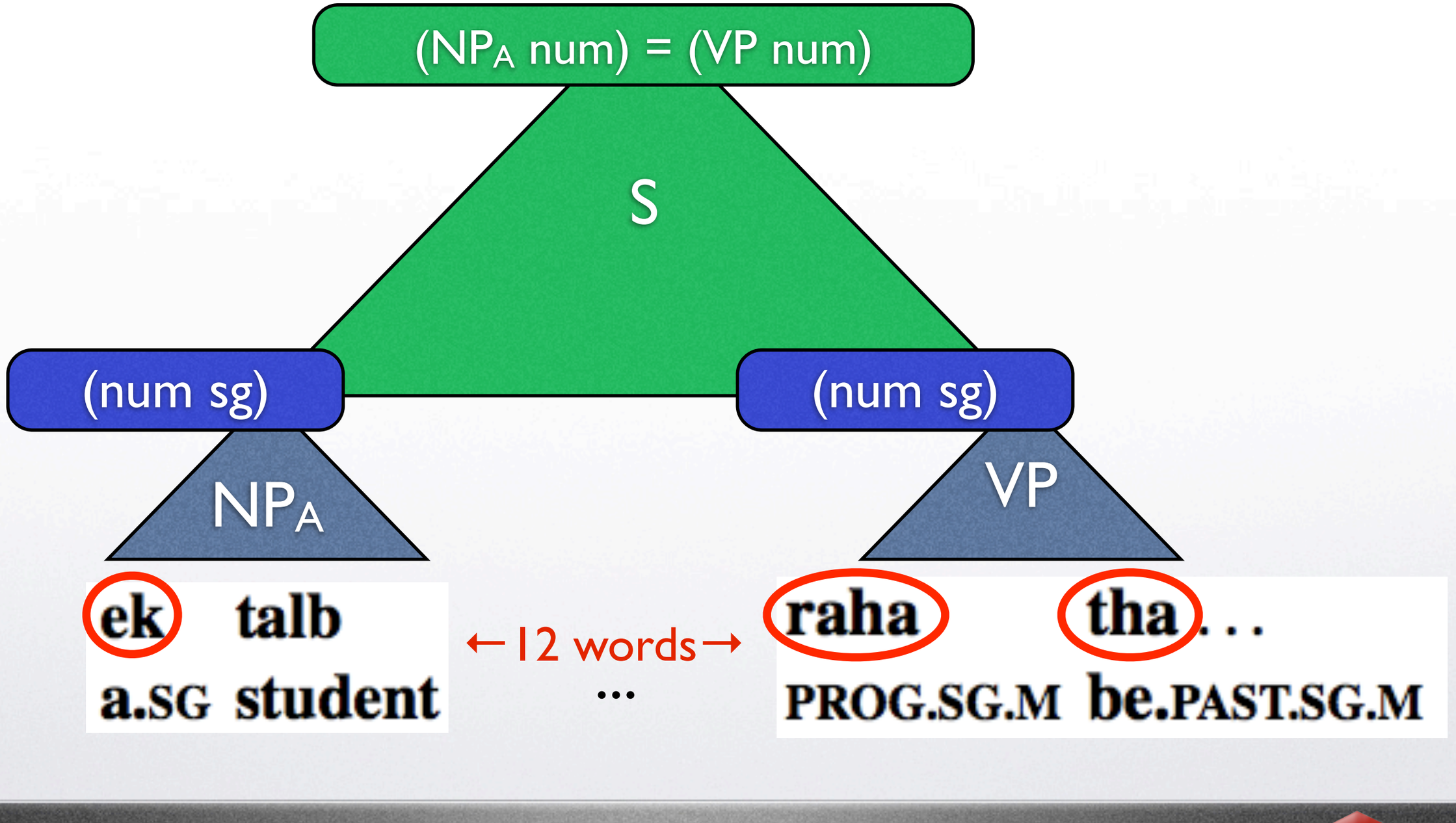
Example Application: Feature-Rich Grammars



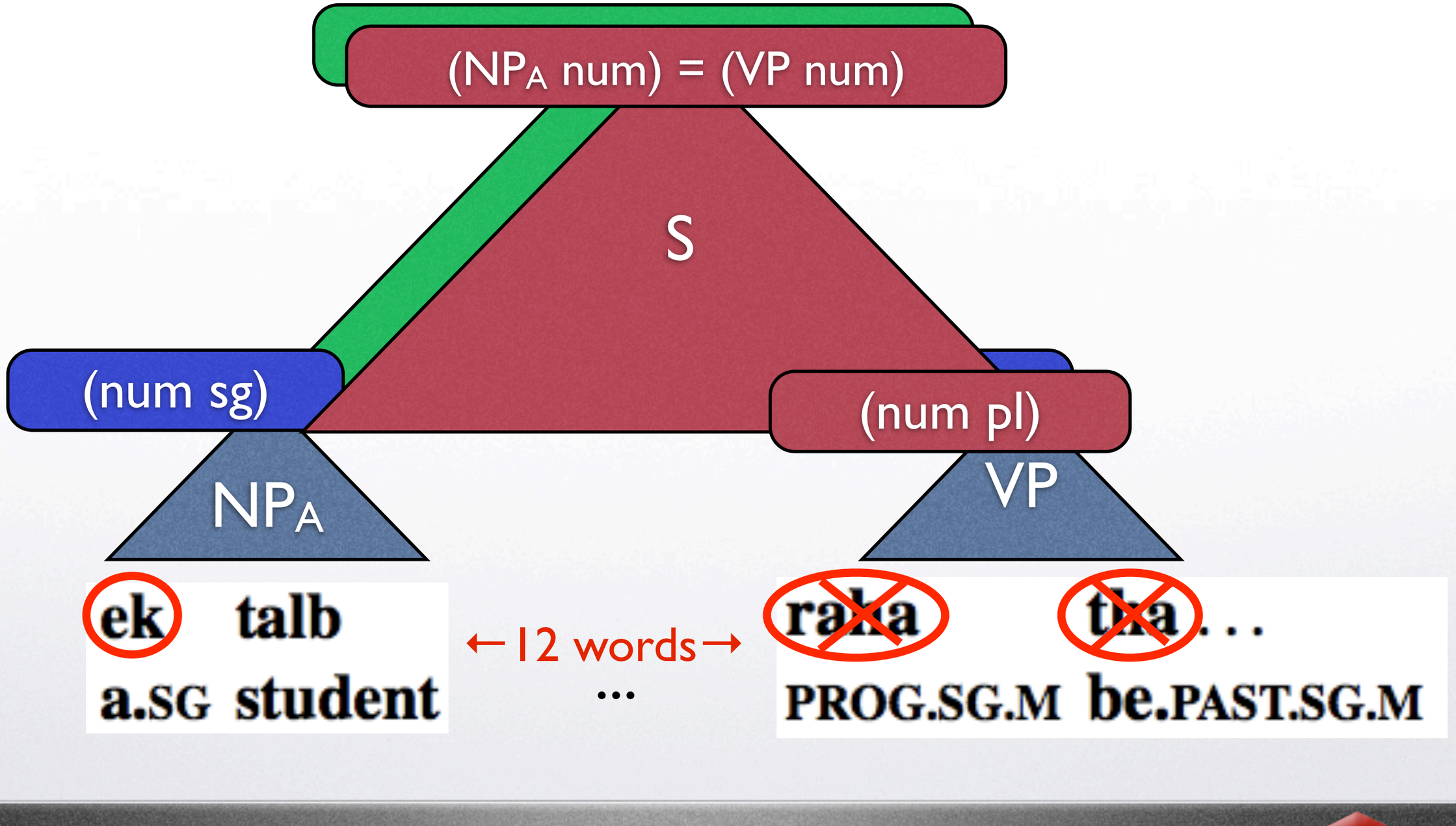
Example Application: Feature-Rich Grammars



Example Application: Feature-Rich Grammars



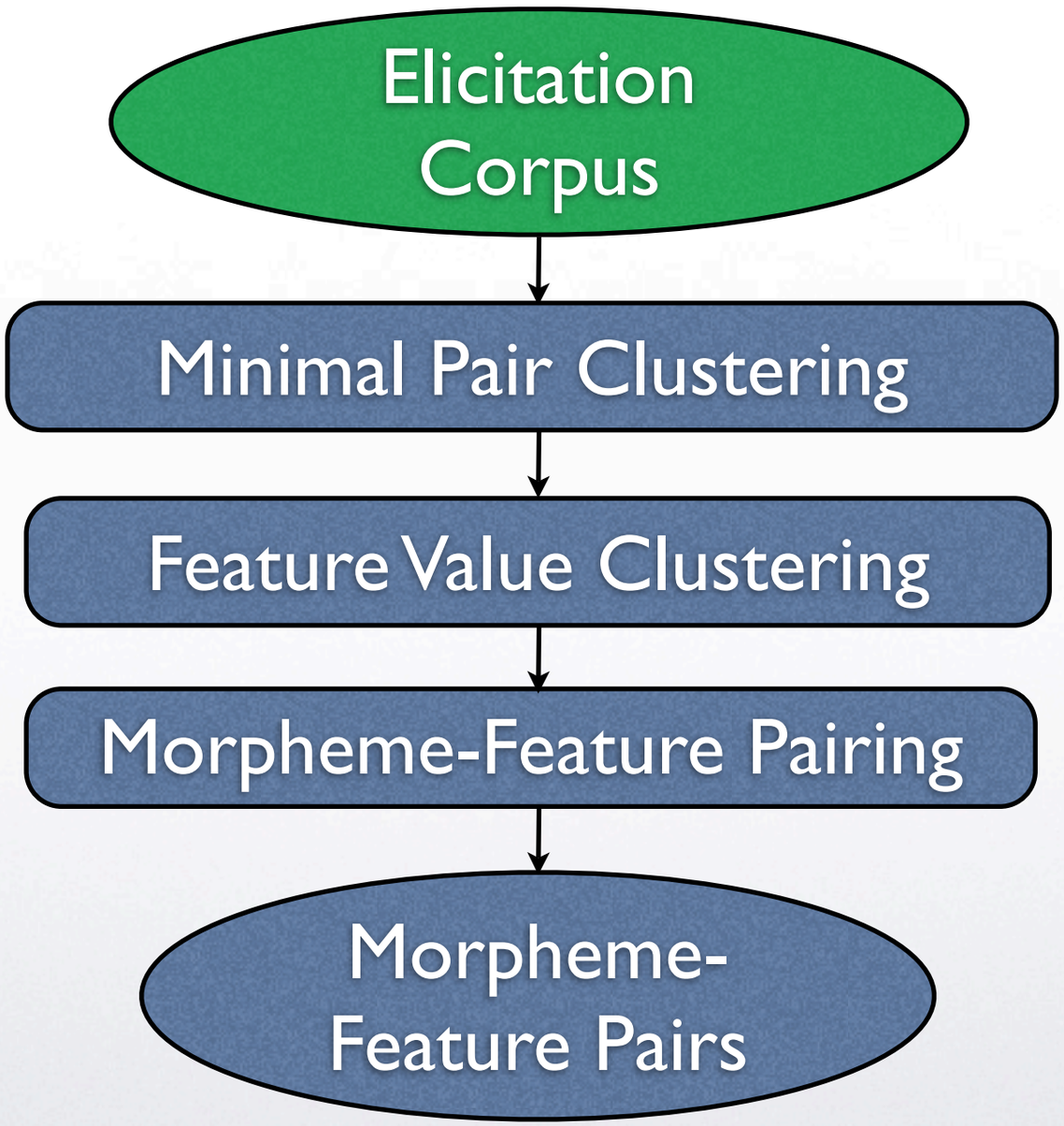
Example Application: Feature-Rich Grammars



Outline

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

Inductive Feature Detection



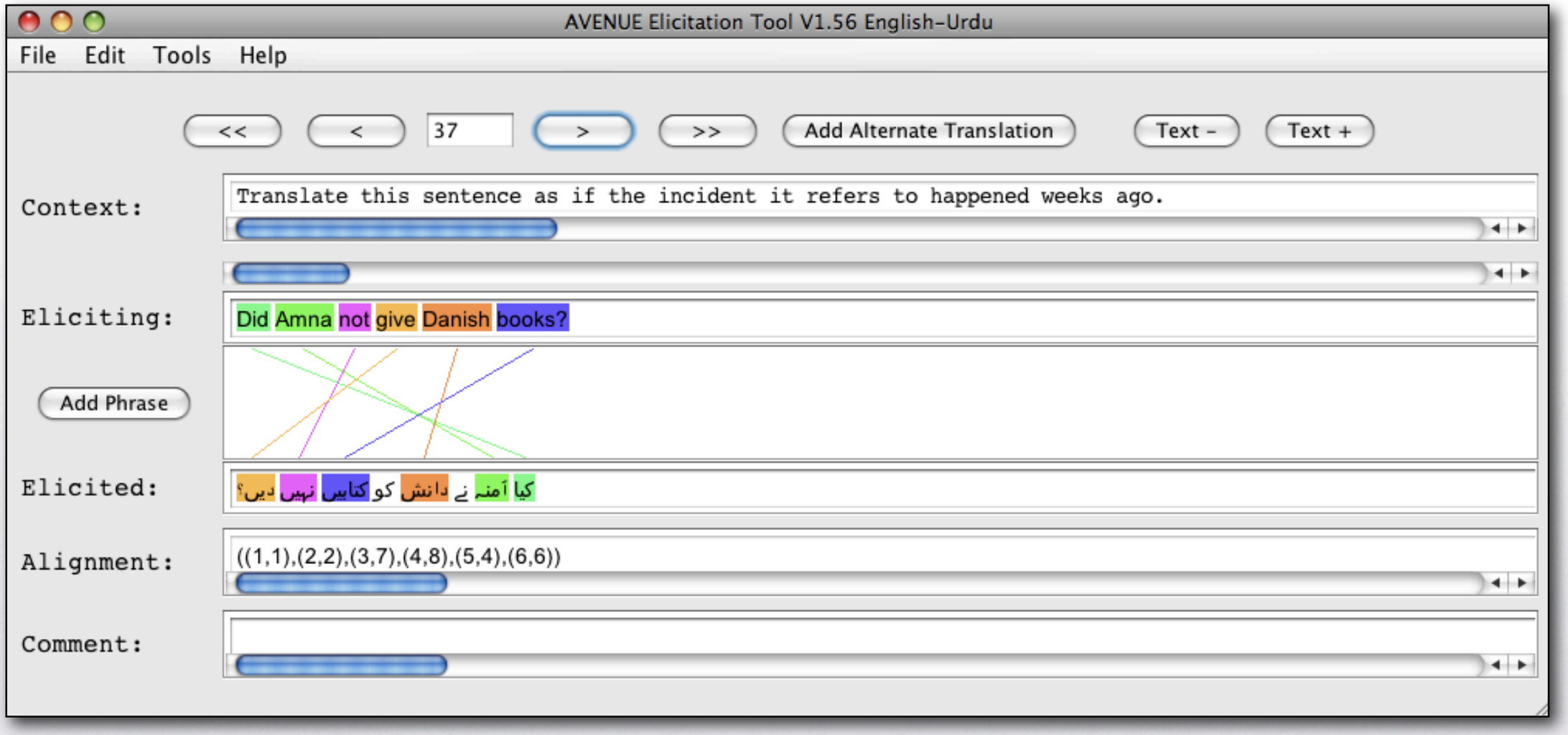
Elicitation Corpus Entry

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 .



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

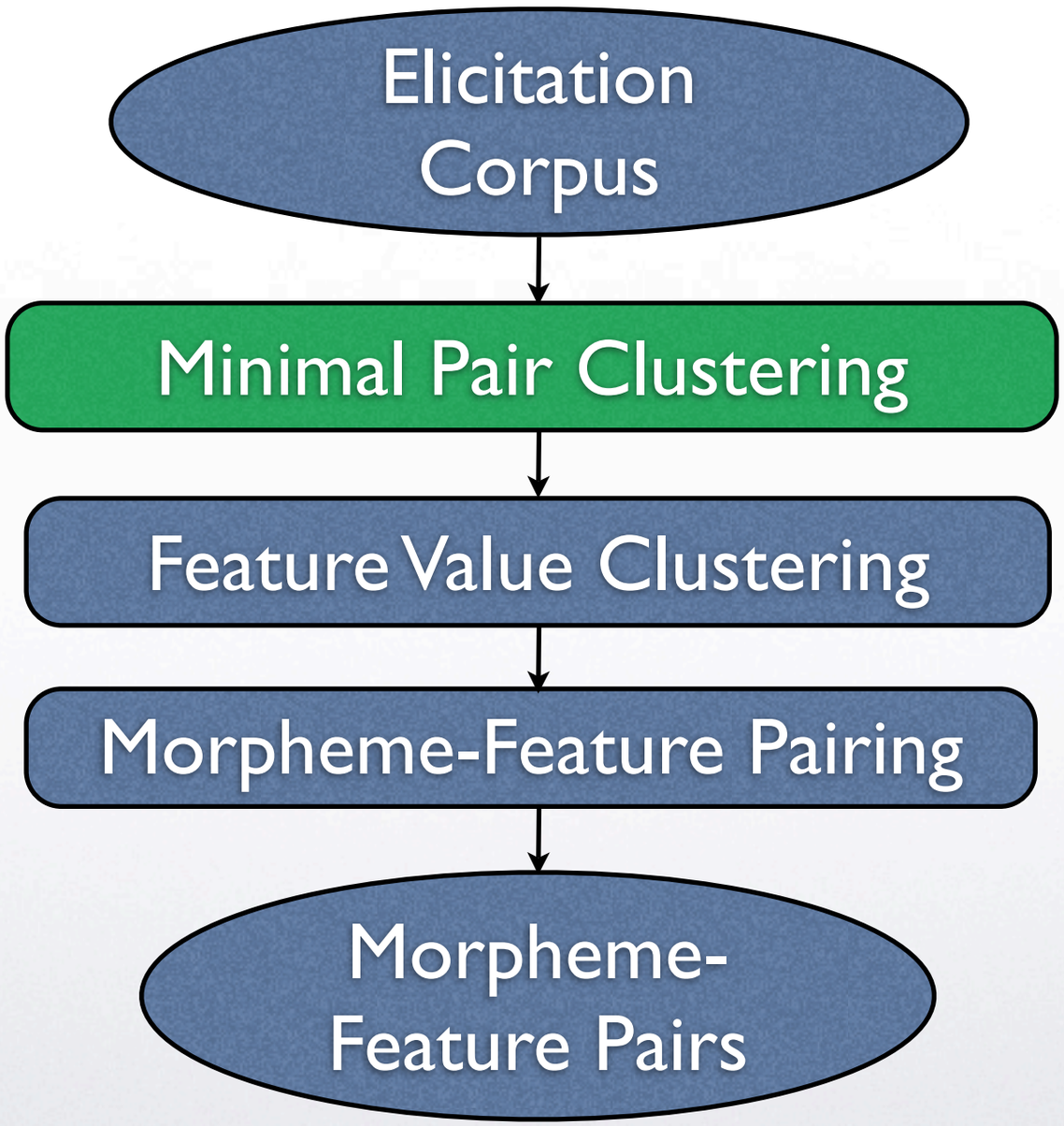
context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies .
tgtsent: Maria hornea galletas .
aligned: ((1,1),(2,2),(3,3),(4,4))
fstruct: [f1]([f2](actor ((gender f)(anim human)(num sg)))
[f3](undergoer ((person 3) (num dl))) (tense pres))

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

The dog sees the cat
El perro ve el gato
((actor (num sg)(gender n))
 (undergoer (num sg)(gender n))
 (tense pres))

pl

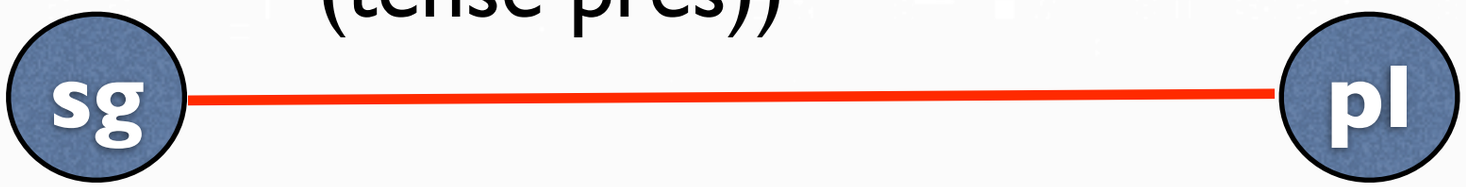
The dogs see the cat
Los perros ven el gato
((actor (num pl)(gender n))
 (undergoer (num sg)(gender n))
 (tense pres))

lti

Minimal Pair Clustering

Wildcard Feature Structure:

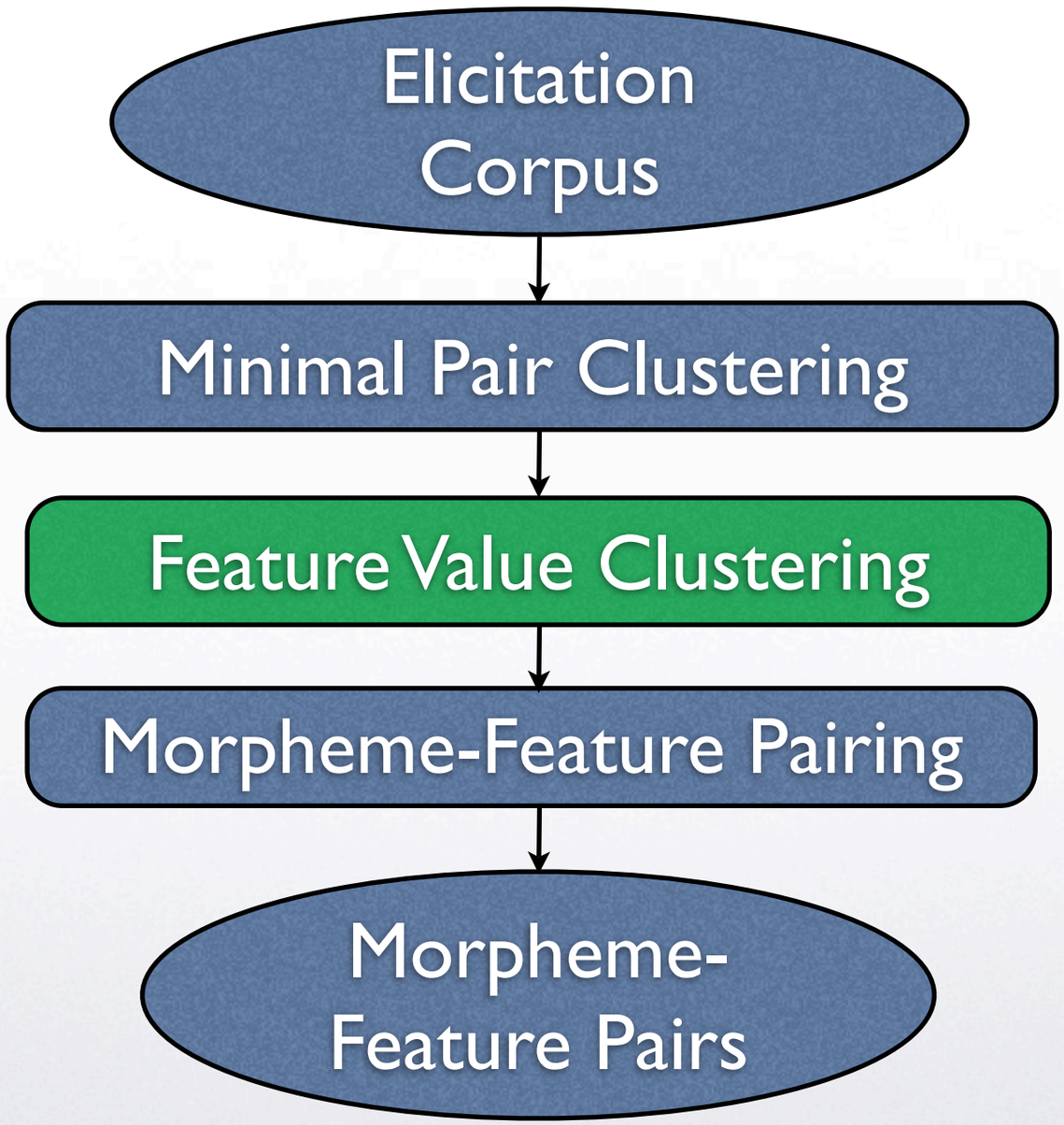
((actor (num *)(gender n))
 (undergoer (num sg)(gender n))
 (tense pres))



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

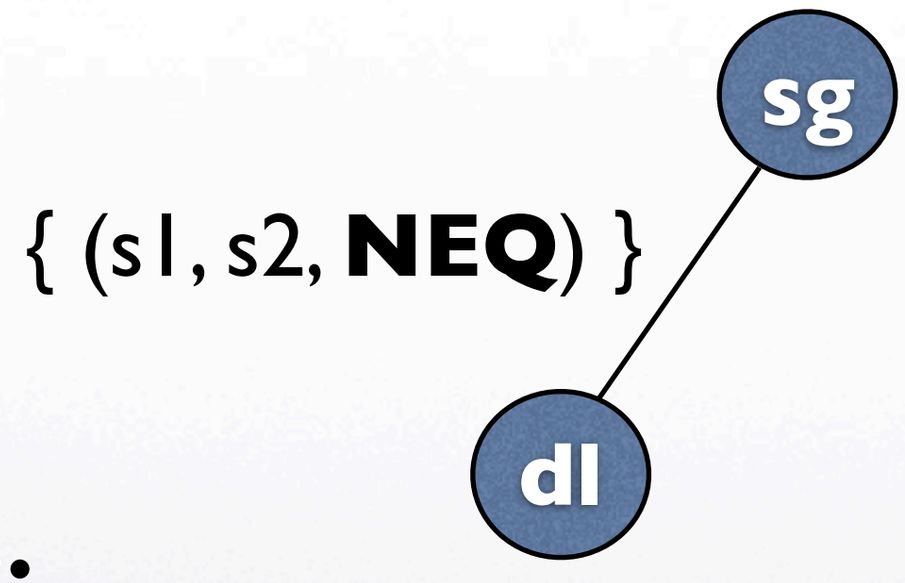
s1: The dog sees the cat

El perro ve el gato

((actor (num sg)…))

Feature Value Clustering

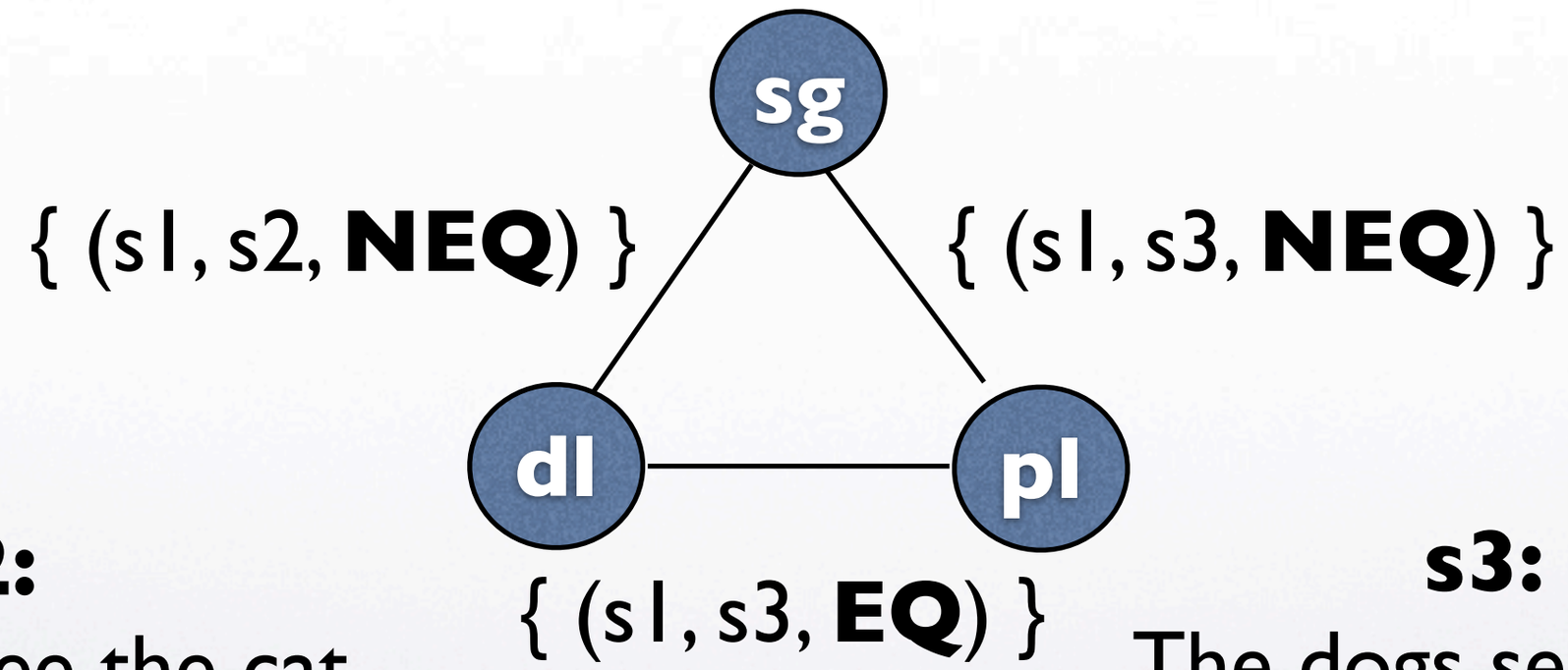
s1: The dog sees the cat
El perro ve el gato
((actor (num sg)…))



s2:
The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

Feature Value Clustering

s1: The dog sees the cat
El perro ve el gato
((actor (num sg)…))



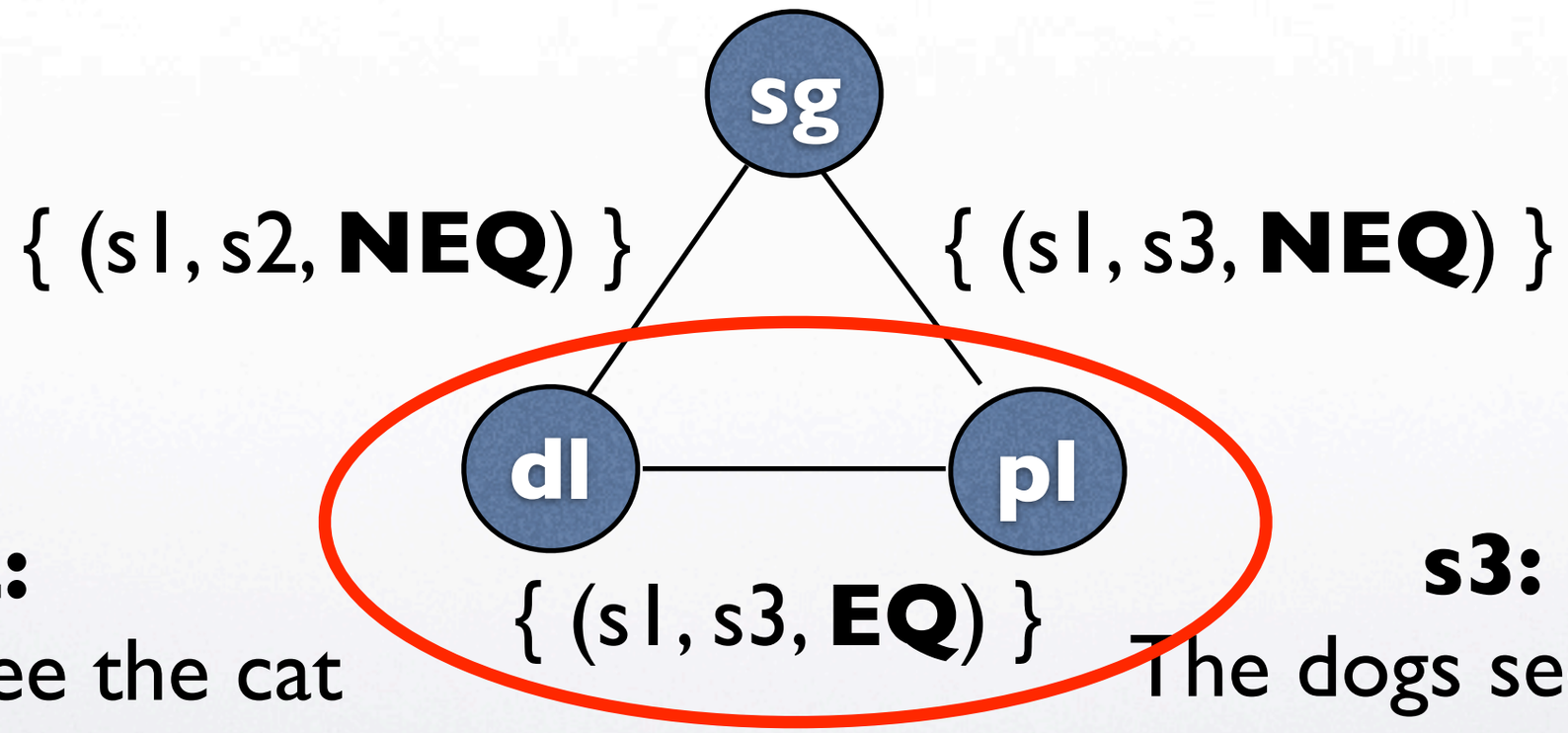
s2:
The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

s3:
The dogs see the cat
Los perros ven el gato
((actor (num pl)…))

lti

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

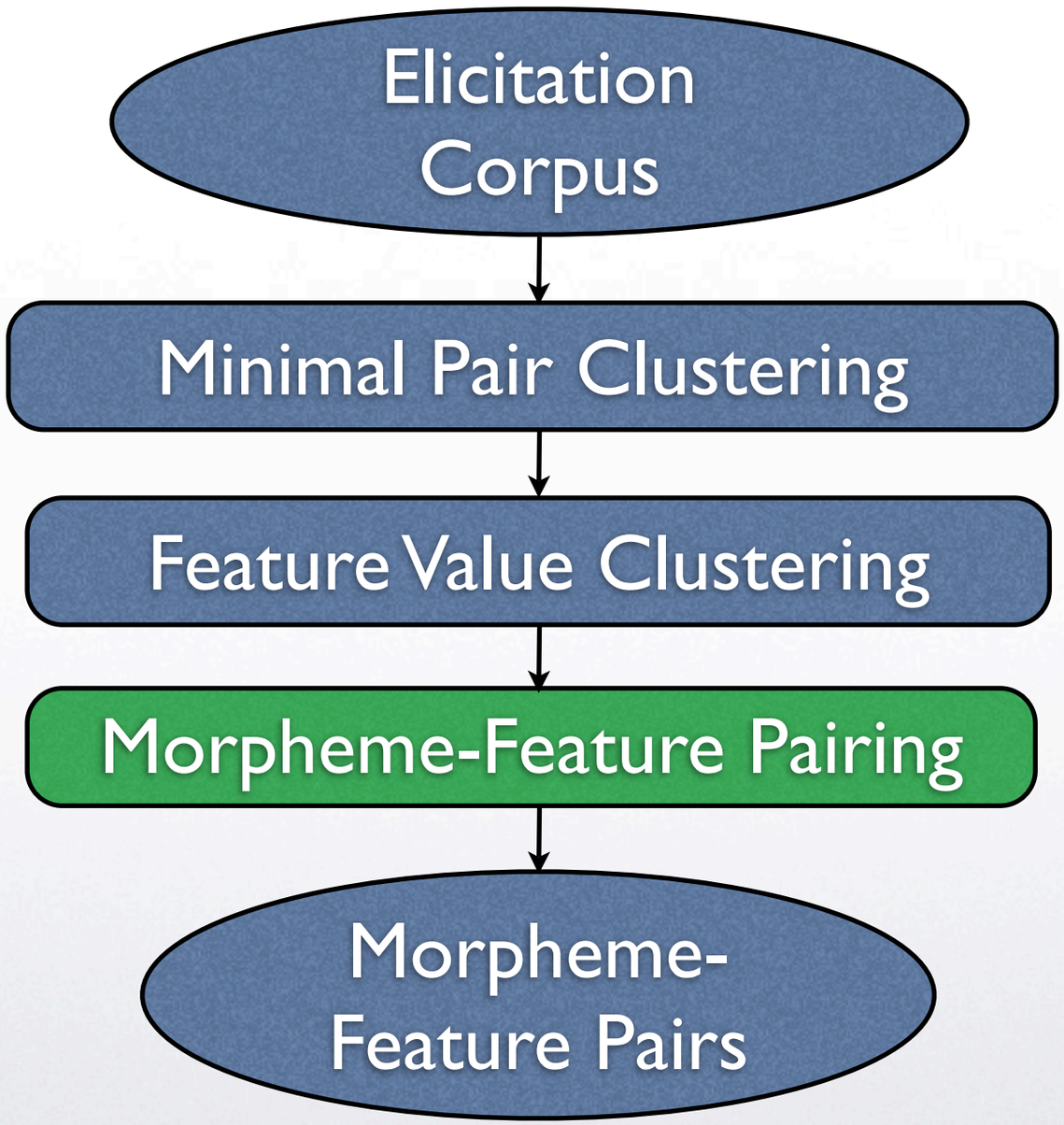
s1: The dog sees the cat
El perro ve el gato
((actor (num sg)…))

s2: The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

{ (s1, s2, **NEQ**),
(s1, s3, **NEQ**) }

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)…))

{ (s1, s2, **NEQ**),
(s1, s3, **NEQ**) }

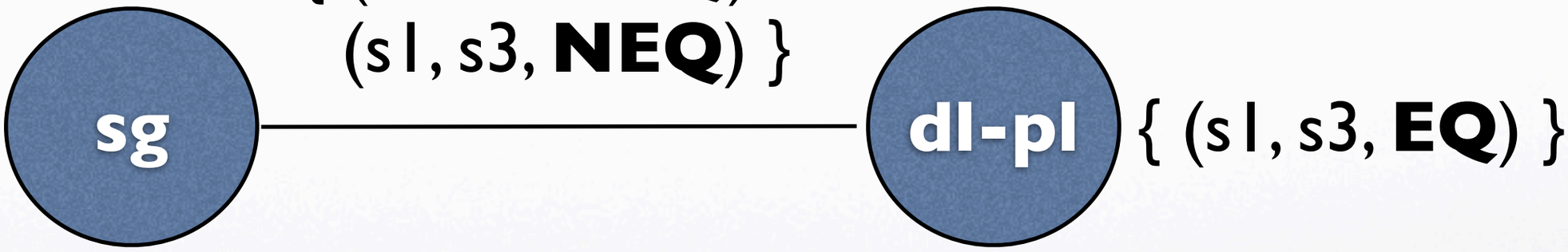
s3: The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

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)…))

{ (s1, s2, **NEQ**),
(s1, s3, **NEQ**) }



{ el, perro, ve }

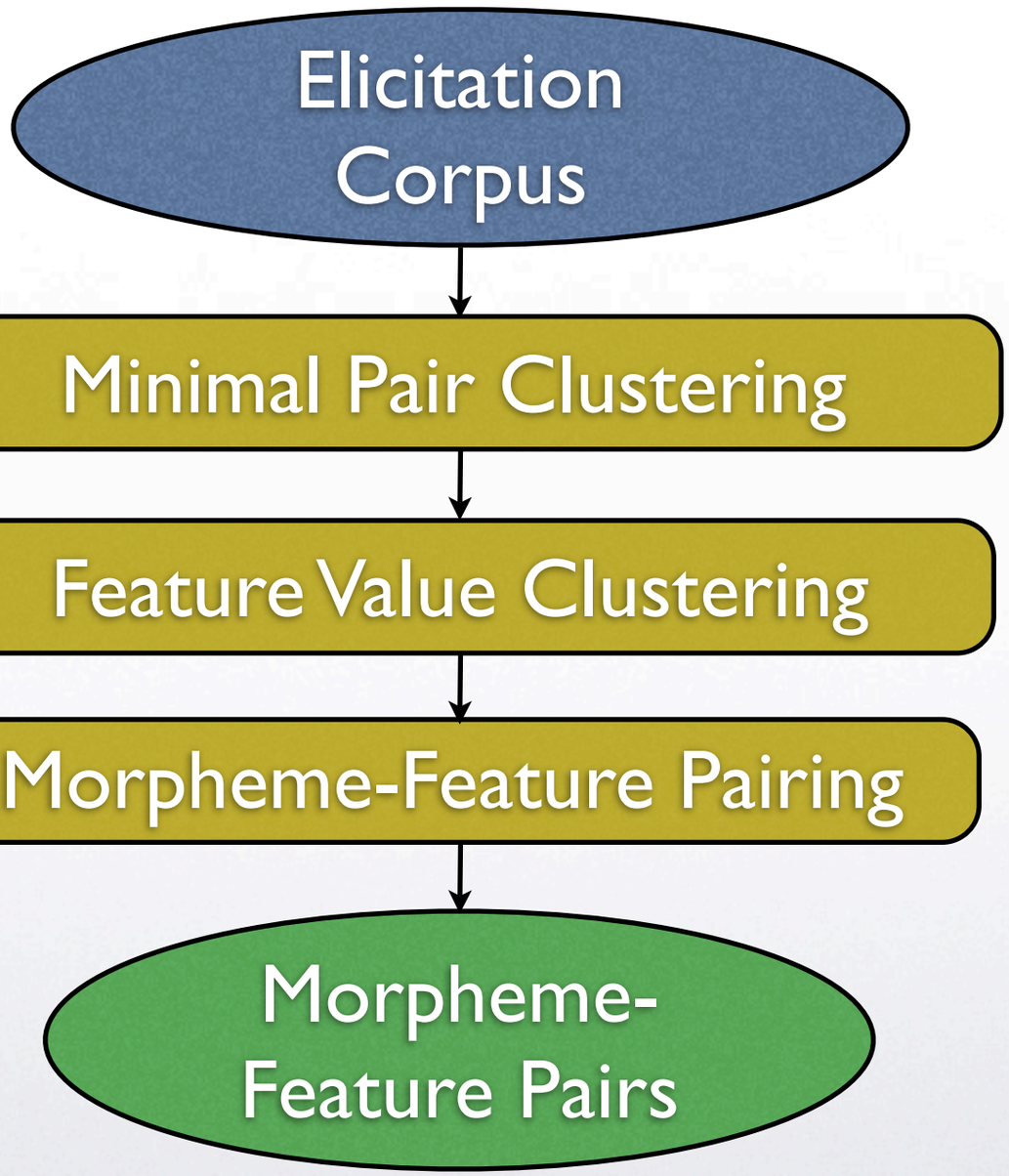
{ los, perros, ven }

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


Elicitation Corpus
Minimal Pair Clustering
Feature Value Clustering
Morpheme-Feature Pairing
Morpheme-Feature Pairs

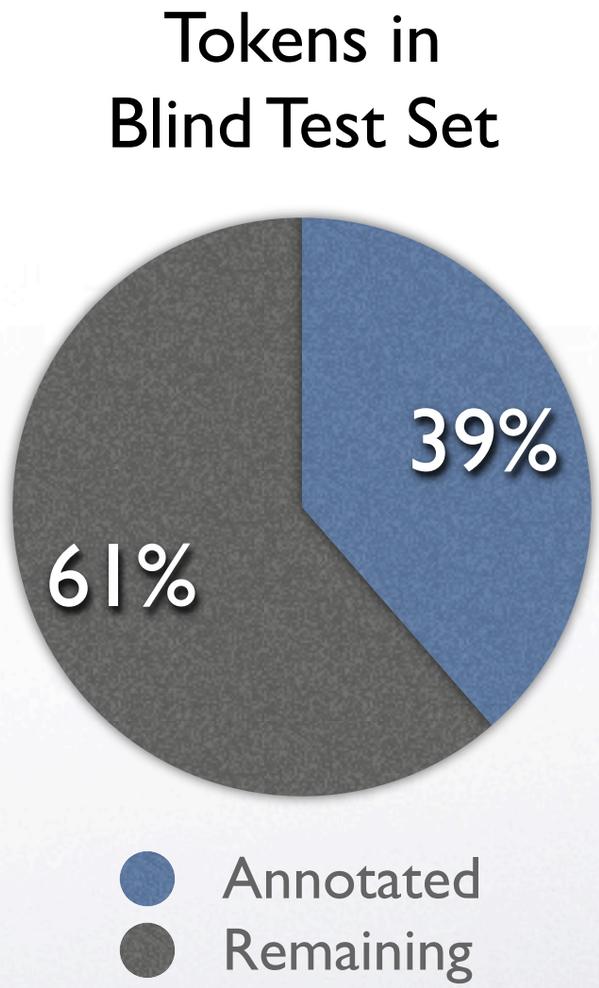
Experiment

- 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	

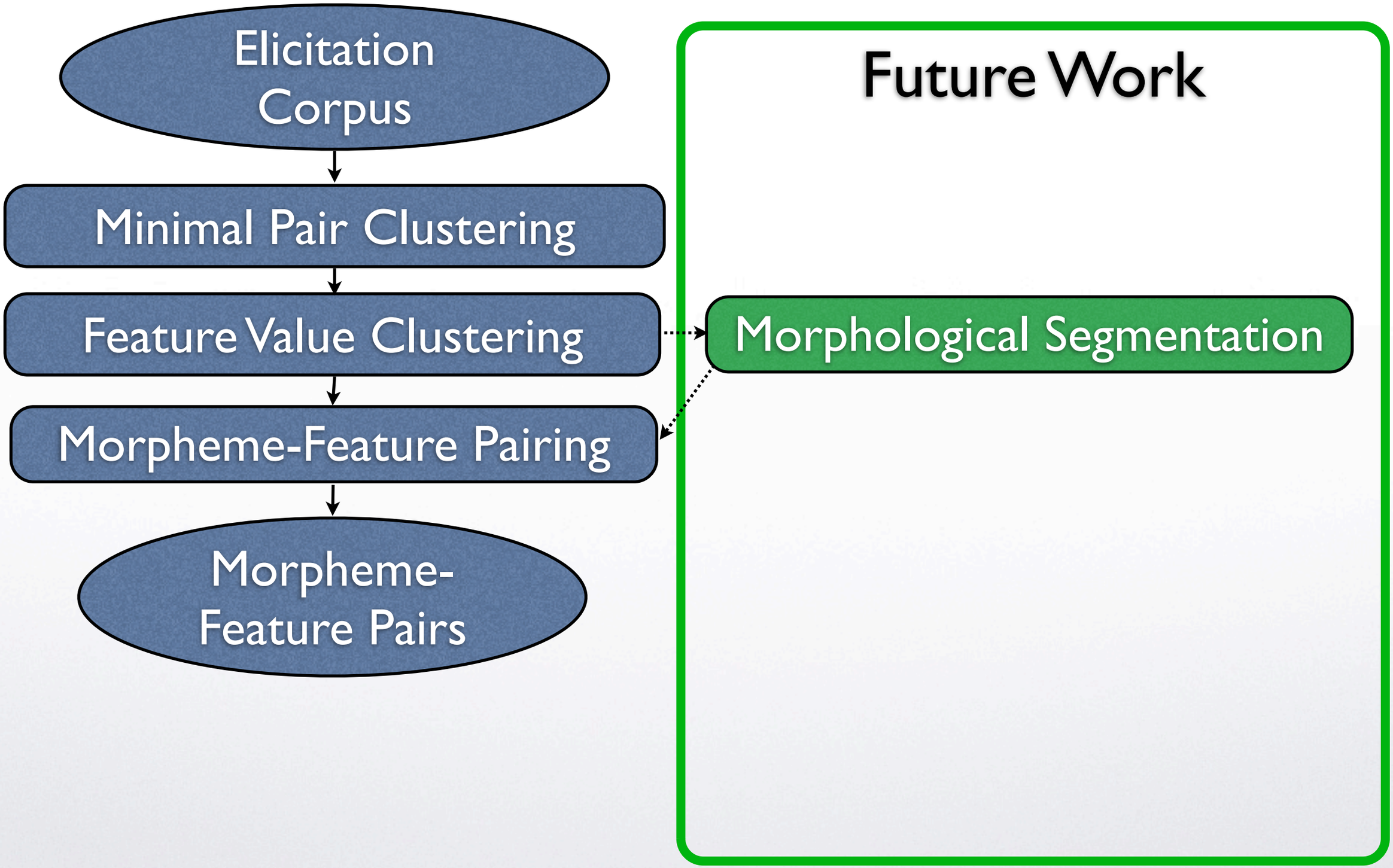
Experiment

- 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



Outline

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


Elicitation Corpus
Minimal Pair Clustering
Feature Value Clustering
Morpheme-Feature Pairing
Morpheme-Feature Pairs
Future Work
Morphological Segmentation

Elicitation Corpus

Minimal Pair Clusterin

Feature Value Clustering

Morpheme-Feature Pai

Morpheme-Feature Pairs

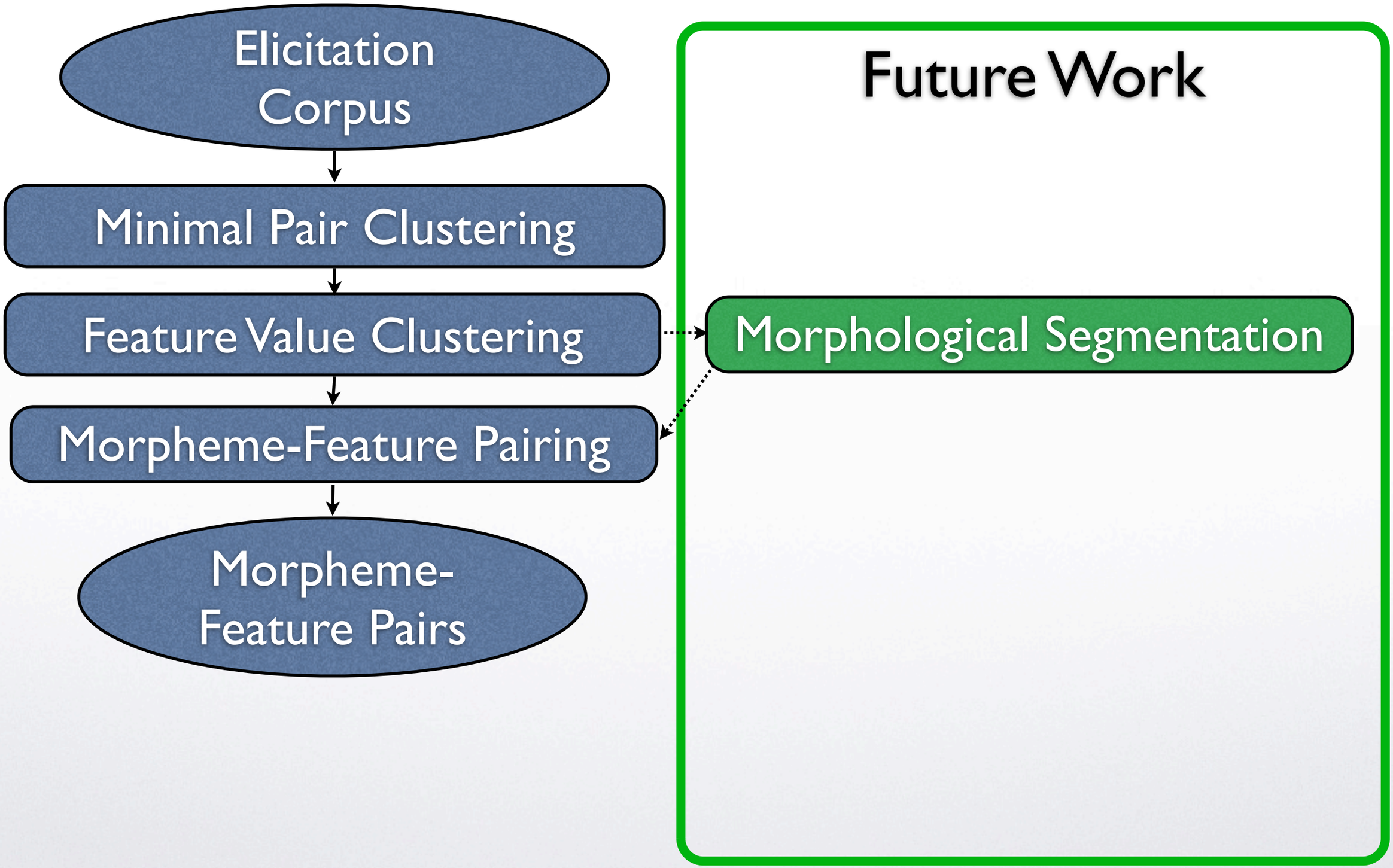
Presented Method

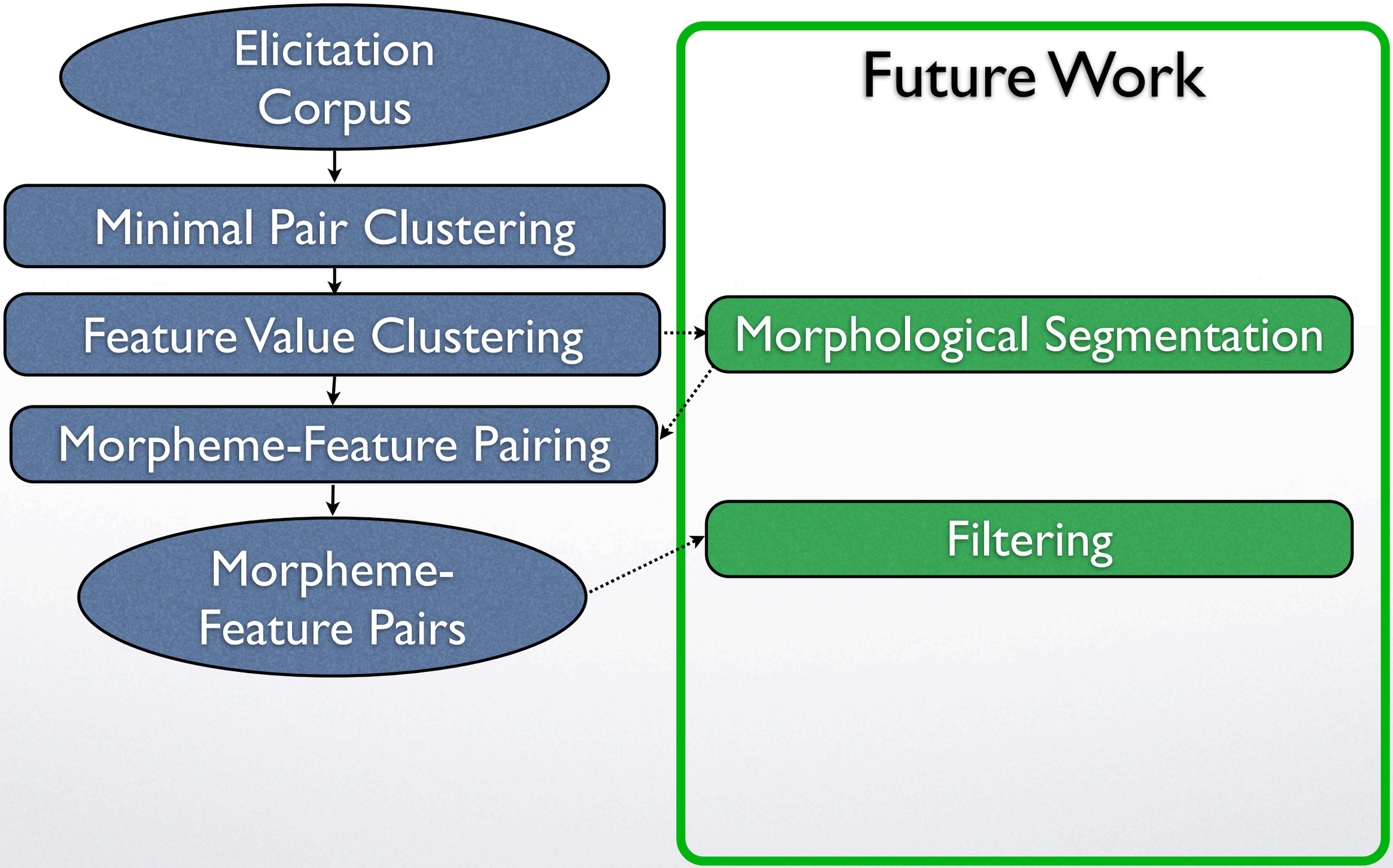
Feature	Candidate Lexical Items
(num sg)	come, el, perro
(num dl, pl)	comen, los, perros

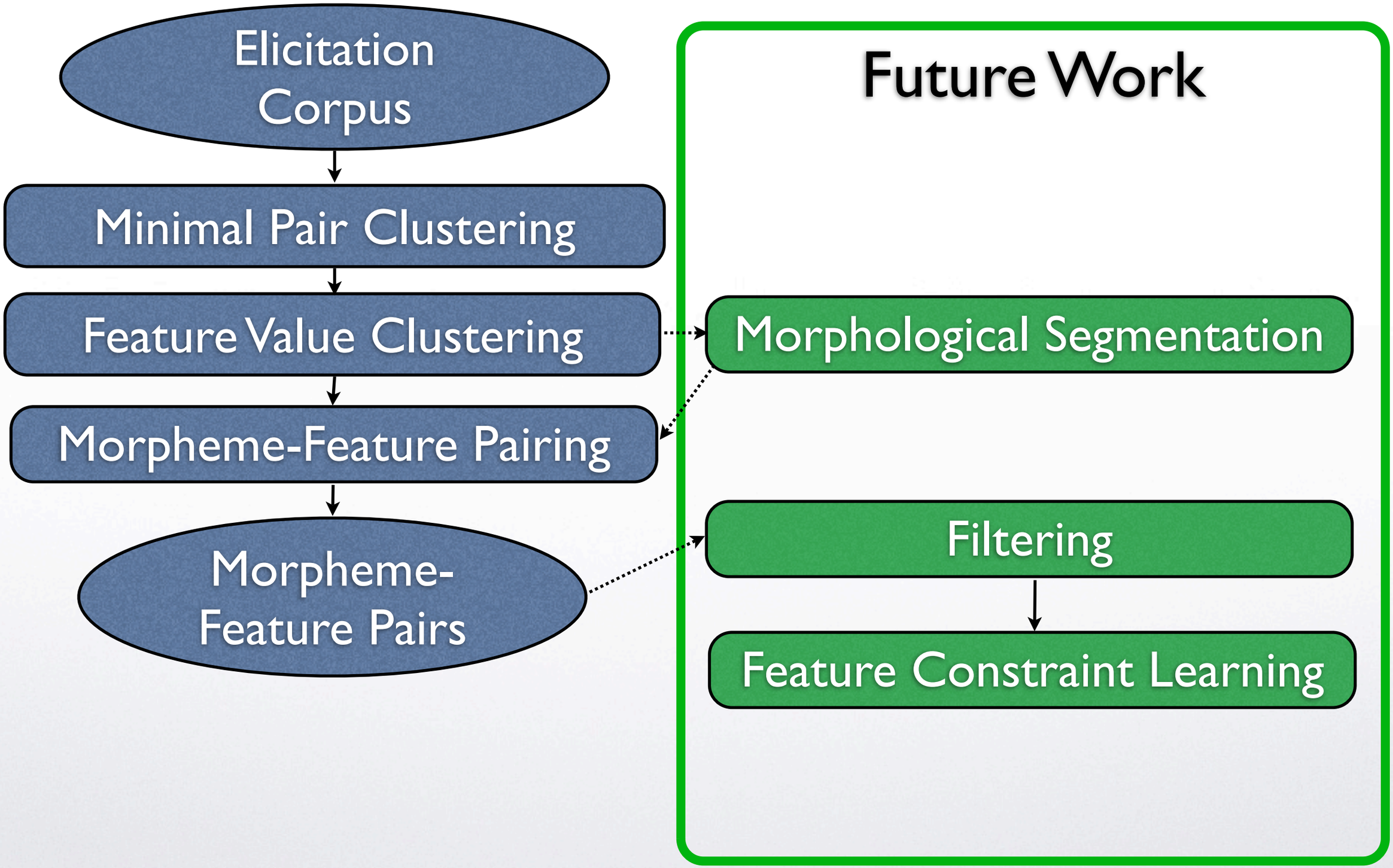
Morphological Segmentation

Future Work

Feature	Candidate Morphemes
(num sg)	VB+Ø, el, N+o
(num dl, pl)	VB+n, los, N+os


Elicitation Corpus
Minimal Pair Clustering
Feature Value Clustering
Morpheme-Feature Pairing
Morpheme-Feature Pairs
Future Work
Morphological Segmentation


Elicitation Corpus
Minimal Pair Clustering
Feature Value Clustering
Morpheme-Feature Pairing
Morpheme-Feature Pairs
Future Work
Morphological Segmentation
Filtering


Elicitation Corpus
Minimal Pair Clustering
Feature Value Clustering
Morpheme-Feature Pairing
Morpheme-Feature Pairs
Future Work
Morphological Segmentation
Filtering
Feature Constraint Learning

Elicitation Corpus

Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Morpheme-Feature Pairs

Future Work

Morphological Segmentation

Filtering

Feature Constraint Learning

Structural MT Integration

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