Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Patrick Xia David Yarowsky

JOHNS HOPKINS UNIVERSITY

The Center for Language and Speech Processing

code, resources, slides: www.github.com/pitrack/monolign

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

I like Taipei

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

I like Taipei
我喜欢台北

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

I like Taipei

我喜欢台北

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

I like Taiwan

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

I like Taiwan

我喜欢台北

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

I like Taiwan

我喜欢台北

我♡台北

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

I like Taiwan

我喜欢台北

我♡台北

I love Taipei

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

I like Taiwan

我喜欢台北

我♡台北

I love Taipei

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

N:1sg V N:Loc

I like Taiwan

我喜欢台北

我♡台北

I love Taipei

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

Multi-parallel corpora

N:1sg V N:Loc

I like Taiwan

我喜欢台北

我♡台北

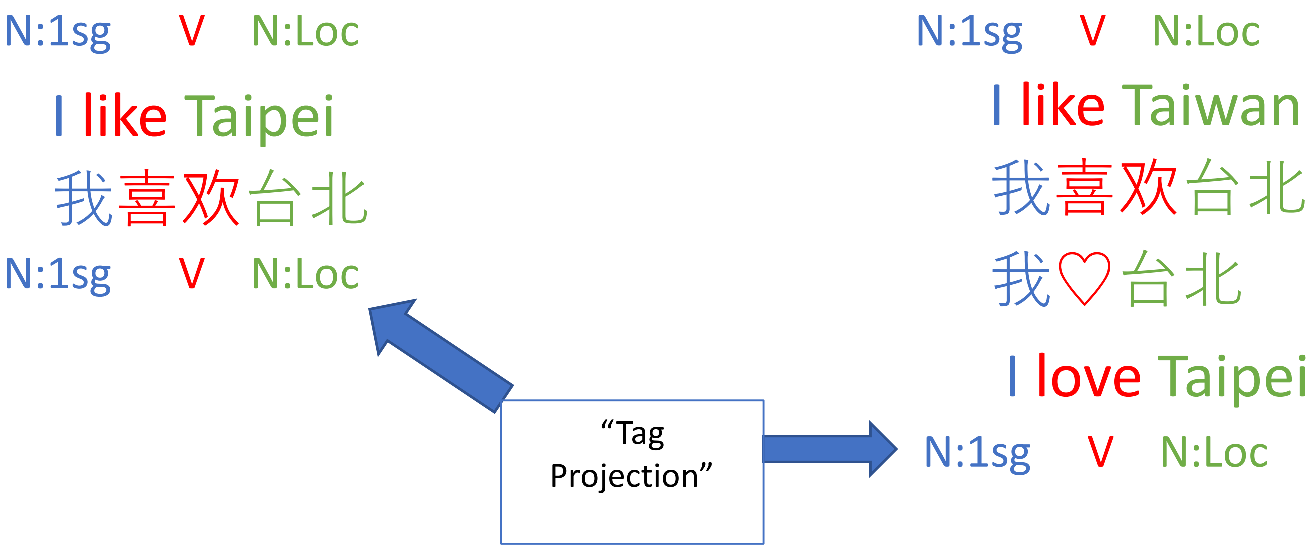
I love Taipei

N:1sg V N:Loc

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Bitext

Multi-parallel corpora



N:1sg V N:Loc

I like Taipei

我喜欢台北

N:1sg V N:Loc

N:1sg V N:Loc

I like Taiwan

我喜欢台北

我♡台北

I love Taipei

N:1sg V N:Loc

“Tag Projection”

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

(Mayer and Cysouw, 2014)

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

- 800+ languages, *verse*-aligned
  - 27 English versions (New Testament)

(Mayer and Cysouw, 2014)

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

- 800+ languages, *verse*-aligned
  - 27 English versions (New Testament)
- Improved English consensus → improved low-resource tag projection from a common source

(Mayer and Cysouw, 2014)

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

- 800+ languages, *verse*-aligned
  - 27 English versions (New Testament)
- Improved English consensus → improved low-resource tag projection from a common source

Examples:

New Simplified:	Then Herod secretly called the astrologers …
Montgomery:	Thereupon Herod sent secretly for the Magi …
Lexham:	Then Herod secretly summoned the wise men …

(Mayer and Cysouw, 2014)

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

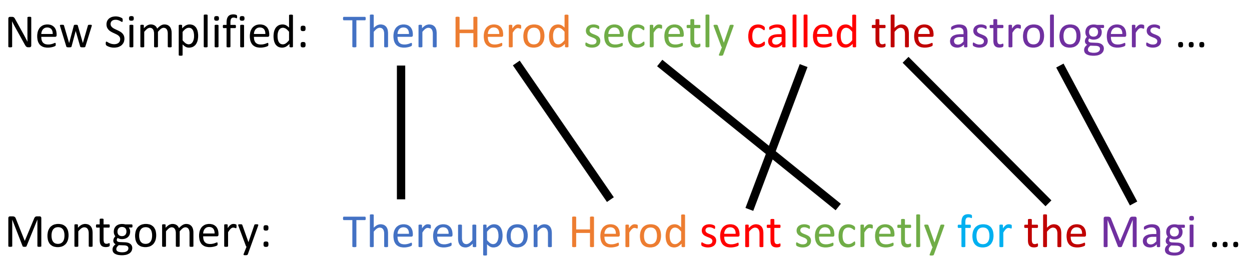
New Simplified: Then Herod secretly called the astrologers …

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

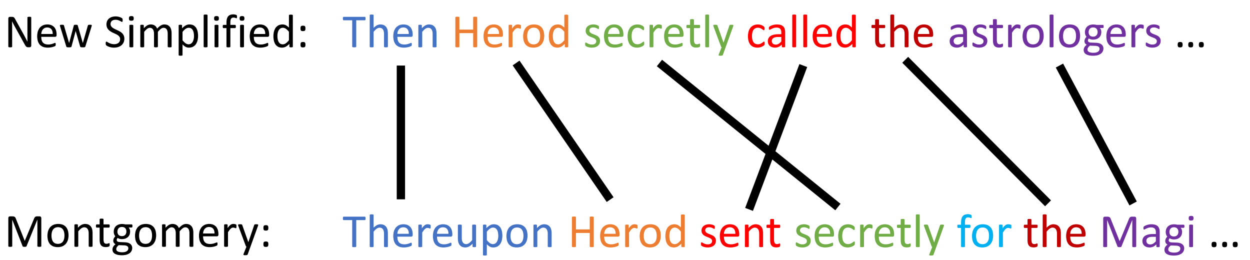
Deriving Consensus for Multi-Parallel Corpora: An English Bible Study



New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

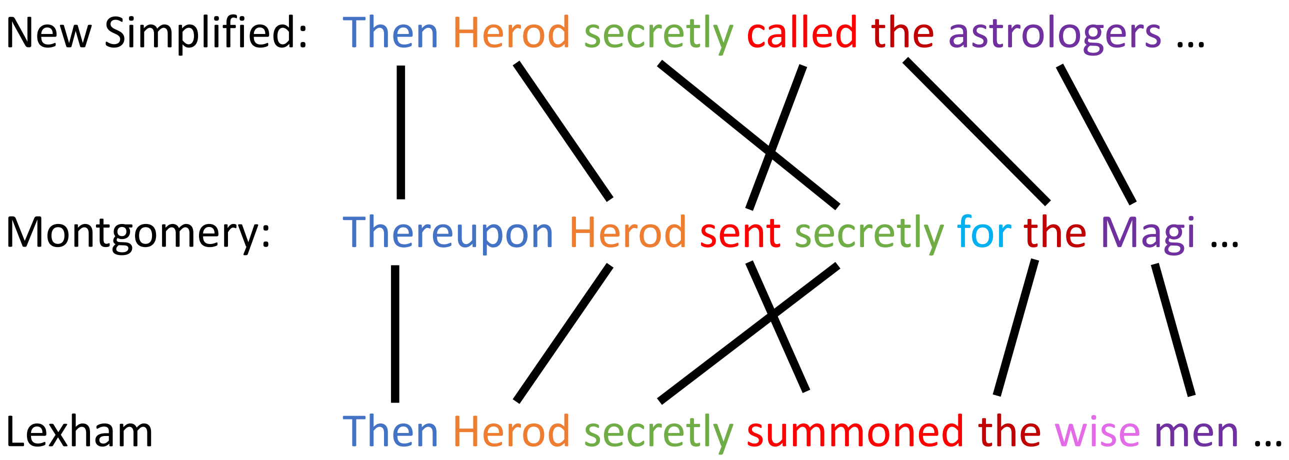


New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

Lexham Then Herod secretly summoned the wise men …

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study

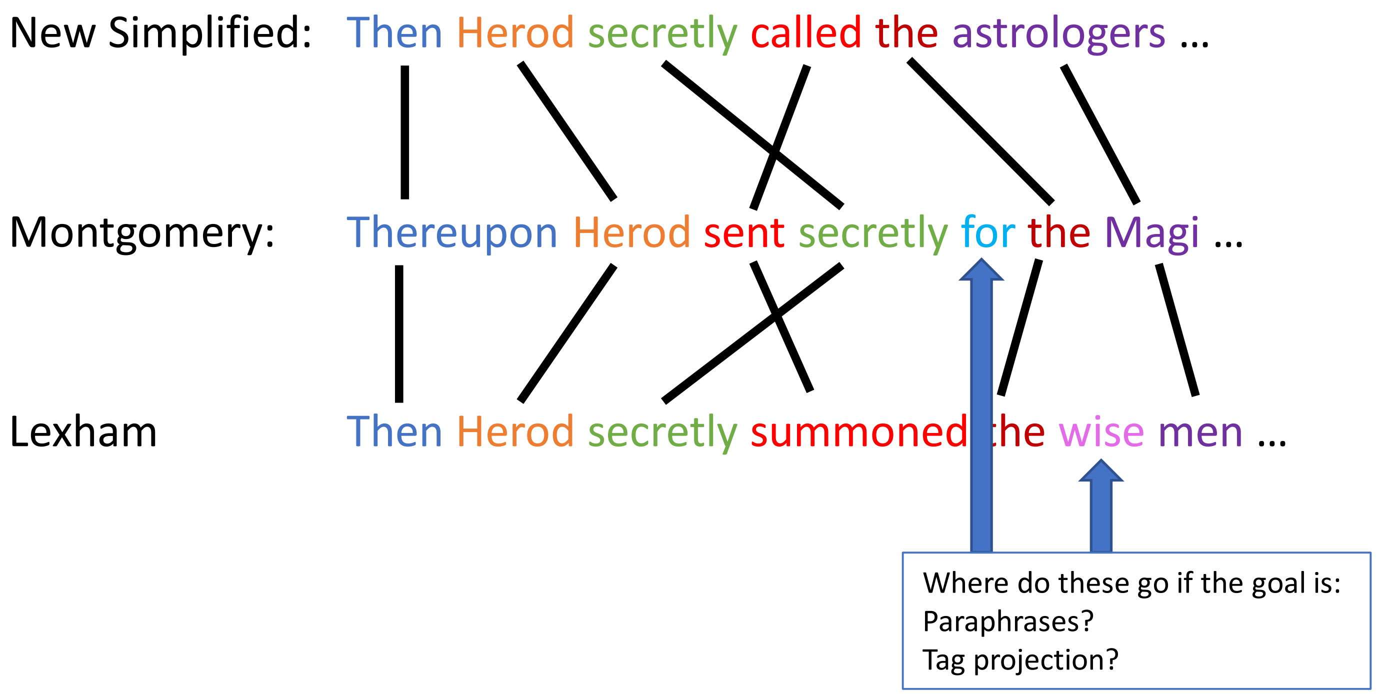


New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

Lexham Then Herod secretly summoned the wise men …

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study



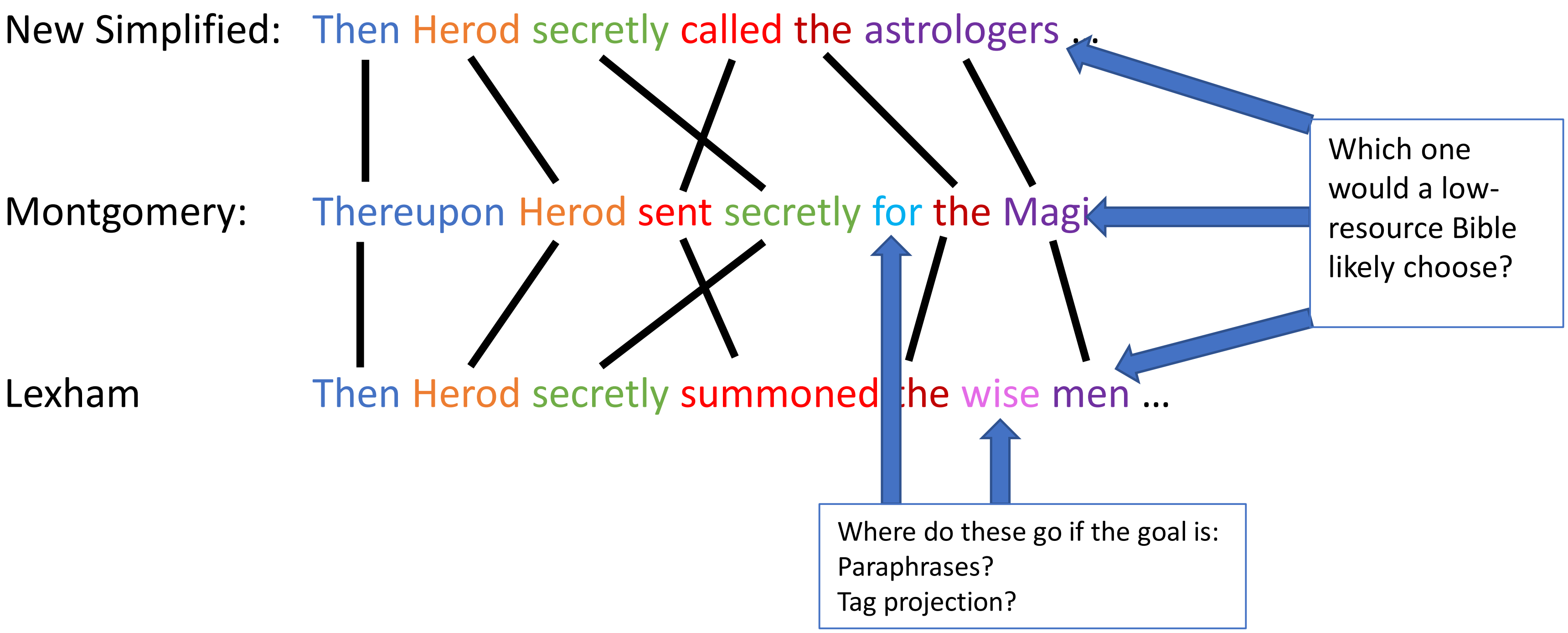
New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

Lexham Then Herod secretly summoned the wise men …

Where do these go if the goal is:
Paraphrases?
Tag projection?

Deriving Consensus for Multi-Parallel Corpora: An English Bible Study



New Simplified: Then Herod secretly called the astrologers ...

Montgomery: Thereupon Herod sent secretly for the Magi

Lexham Then Herod secretly summoned the wise men ...

Which one would a low-resource Bible likely choose?

Where do these go if the goal is:
Paraphrases?
Tag projection?

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matchings
3. Consensus + Resources

New Simplified: Then Herod secretly called the astrologers …

Montgomery: Thereupon Herod sent secretly for the Magi …

Lexham: Then Herod secretly summoned the wise men …

Creating a candidate matching

- Randomly select first document

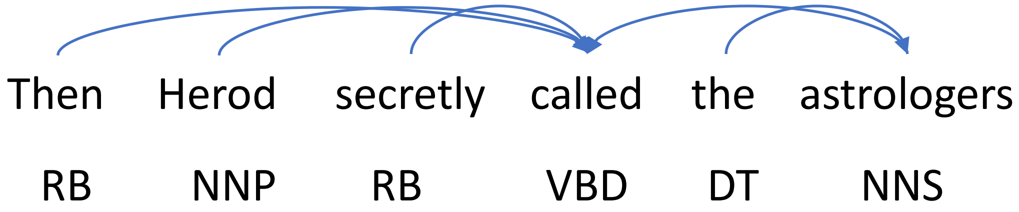
New Simplified:

Then Herod secretly called the astrologers

Creating a candidate matching

- Randomly select first document

New Simplified:



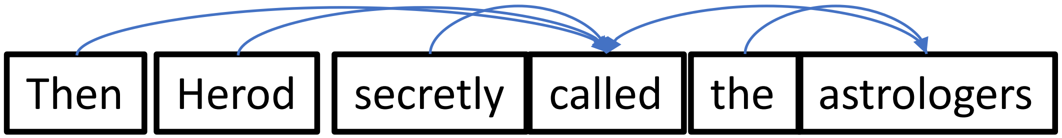
Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Creating a candidate matching

- Initialize matching

Matching:



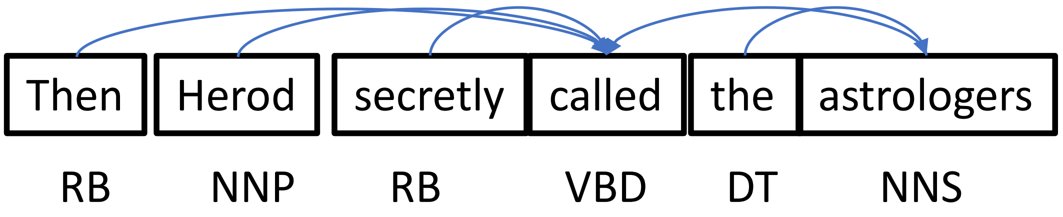
Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Creating a candidate matching

- Initialize matching

Matching:



Then | Herod | secretly | called | the | astrologers

RB | NNP | RB | VBD | DT | NNS

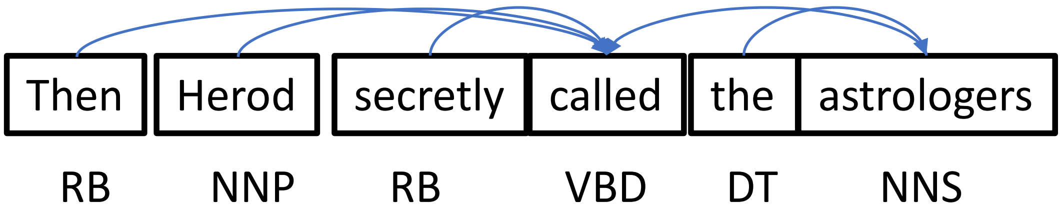
New Simplified

Score: 0

Creating a candidate matching

- For each document, score possible edges

Matching:



Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Montgomery:

Thereupon Herod sent secretly for the Magi

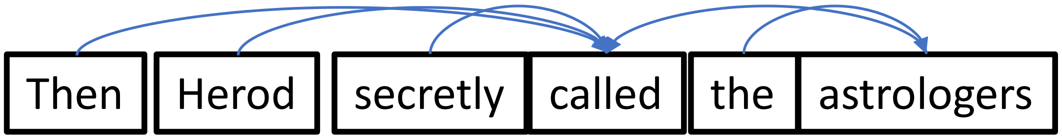
New Simplified
→*Montgomery*

Score: 0

Creating a candidate matching

- For each document, score possible edges

Matching:



Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Montgomery:

RB NNP VBD RB IN DT NNS

Thereupon Herod sent secretly for the Magi

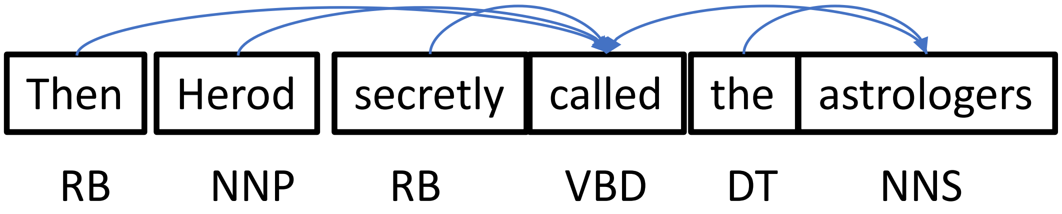
New Simplified
→*Montgomery*

Score: 0

Creating a candidate matching

- For each document, score possible edges

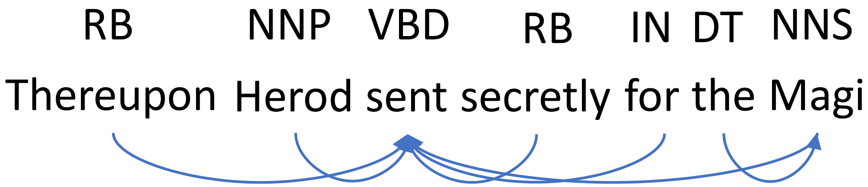
Matching:



Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Montgomery:



RB NNP VBD RB IN DT NNS

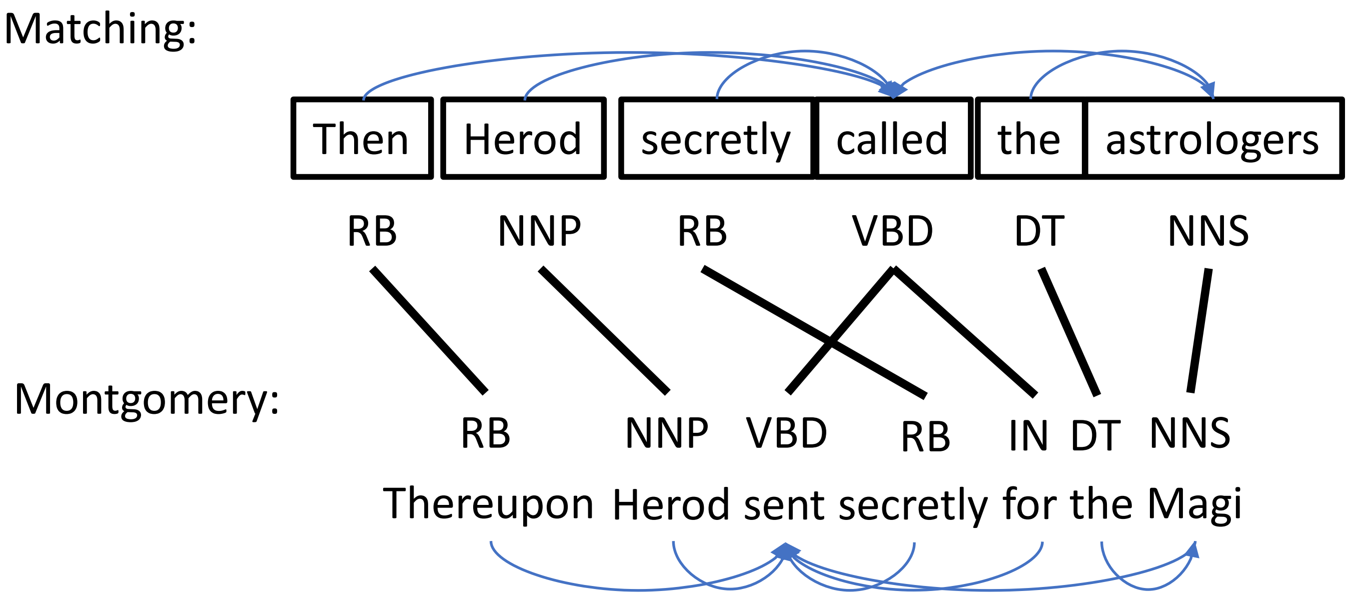
Thereupon Herod sent secretly for the Magi

New Simplified
→*Montgomery*

Score: 0

Creating a candidate matching

- For each document, score possible edges

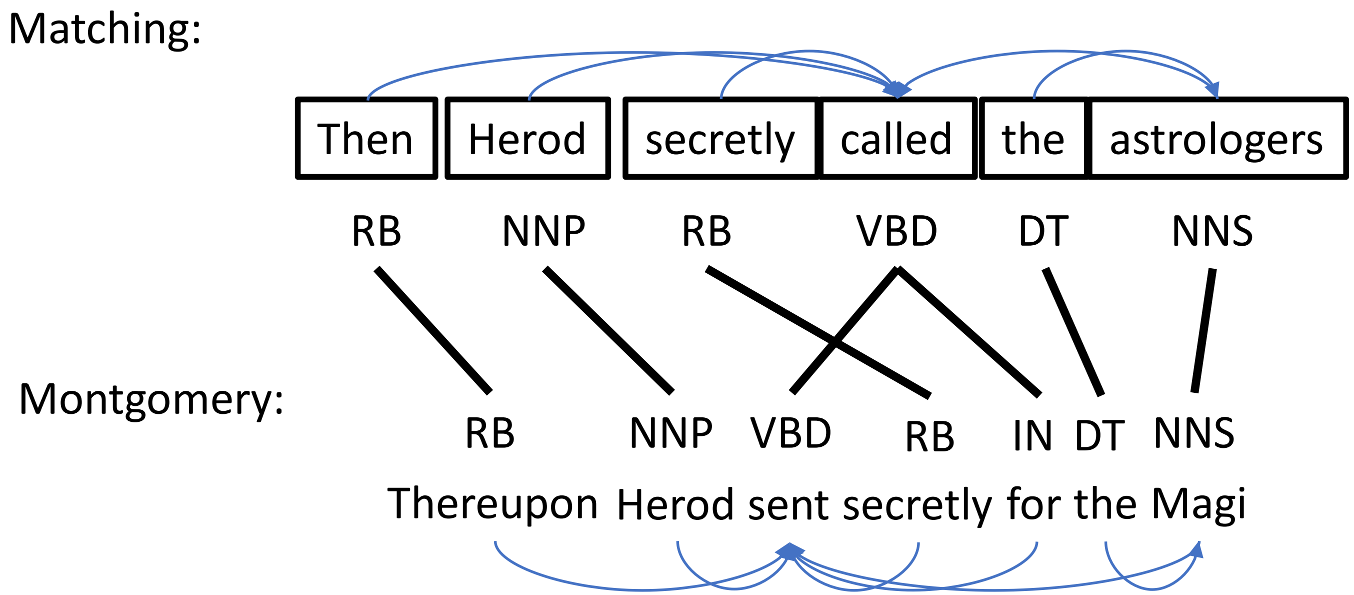


New Simplified →*Montgomery*

Score: 0

Creating a candidate matching

- Take maximum weight matching



Matching:

Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Montgomery:

RB NNP VBD RB IN DT NNS

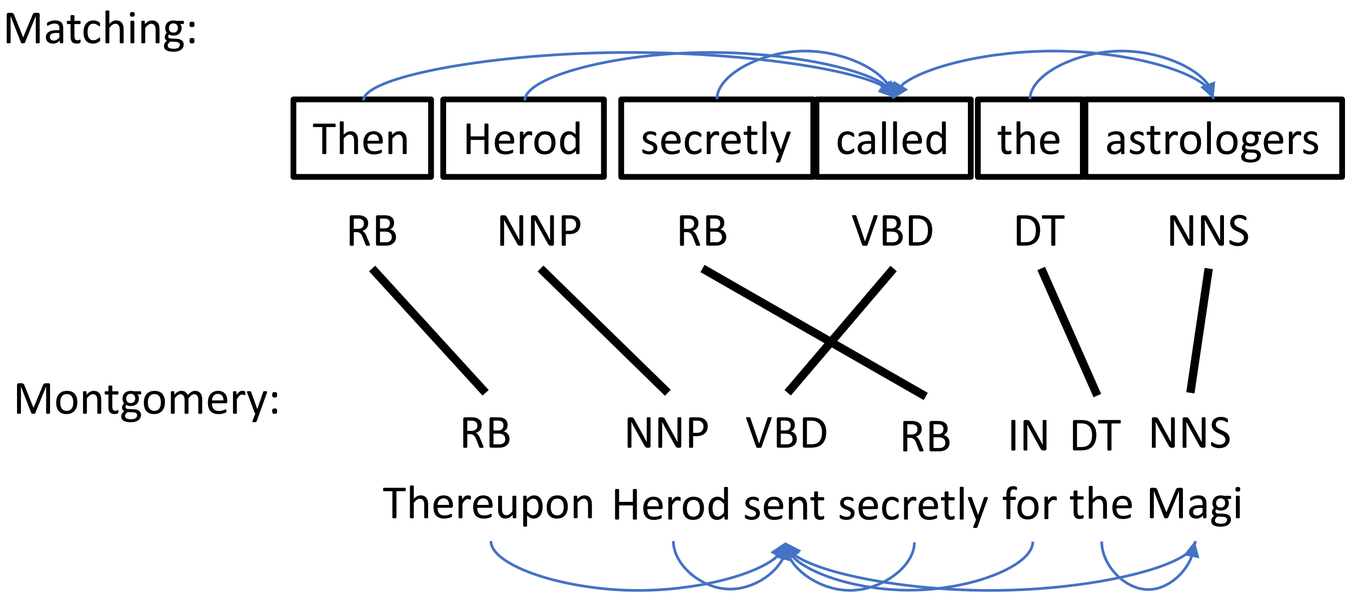
Thereupon Herod sent secretly for the Magi

New Simplified
→*Montgomery*

Score: 0

Creating a candidate matching

- Take maximum weight matching



Matching:

Then Herod secretly called the astrologers

RB NNP RB VBD DT NNS

Montgomery:

RB NNP VBD RB IN DT NNS

Thereupon Herod sent secretly for the Magi

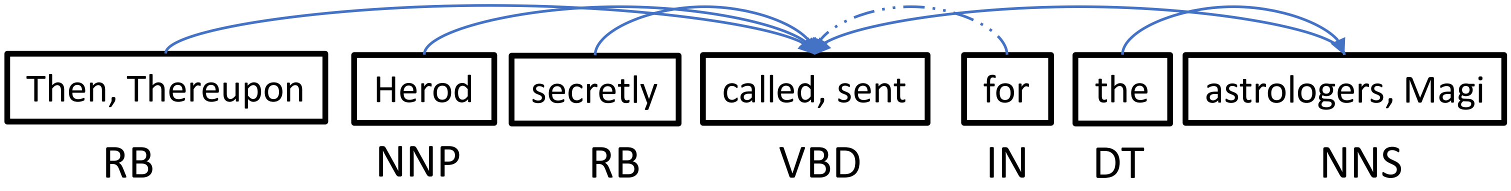
New Simplified
→*Montgomery*

Score: 0

Creating a candidate matching

- Merge

Matching:

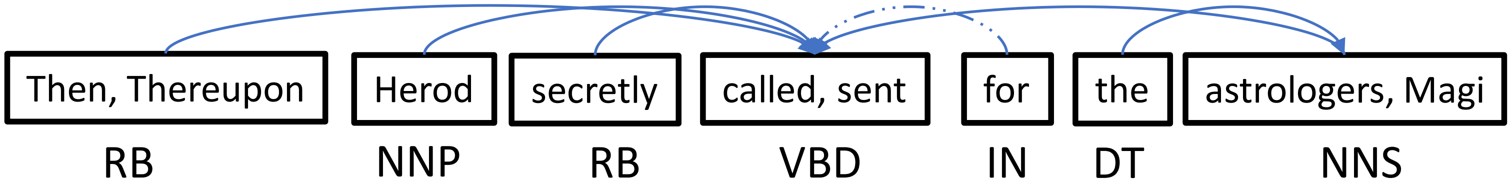


Then, Thereupon	Herod	secretly	called, sent	for	the	astrologers, Magi
RB	NNP	RB	VBD	IN	DT	NNS

Creating a candidate matching

- Merge

Matching:



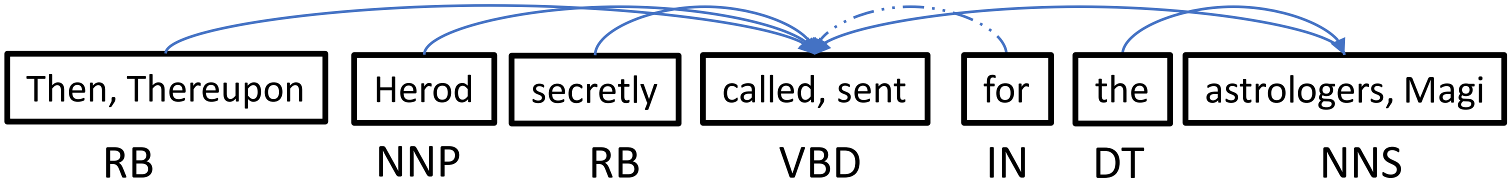
New Simplified Montgomery

Score: 20

Creating a candidate matching

- Repeat for each document

Matching:



Then, Thereupon | Herod | secretly | called, sent | for | the | astrologers, Magi

RB NNP RB VBD IN DT NNS

Lexham:

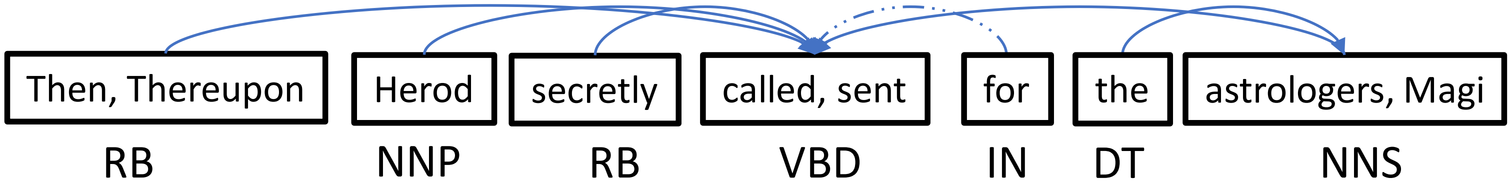
Then Herod secretly summoned the wise men

New Simplified Montgomery
→*Lexham*
Score: 20

Creating a candidate matching

- Repeat for each document

Matching:



Then, Thereupon | Herod | secretly | called, sent | for | the | astrologers, Magi

RB | NNP | RB | VBD | IN | DT | NNS

Lexham:

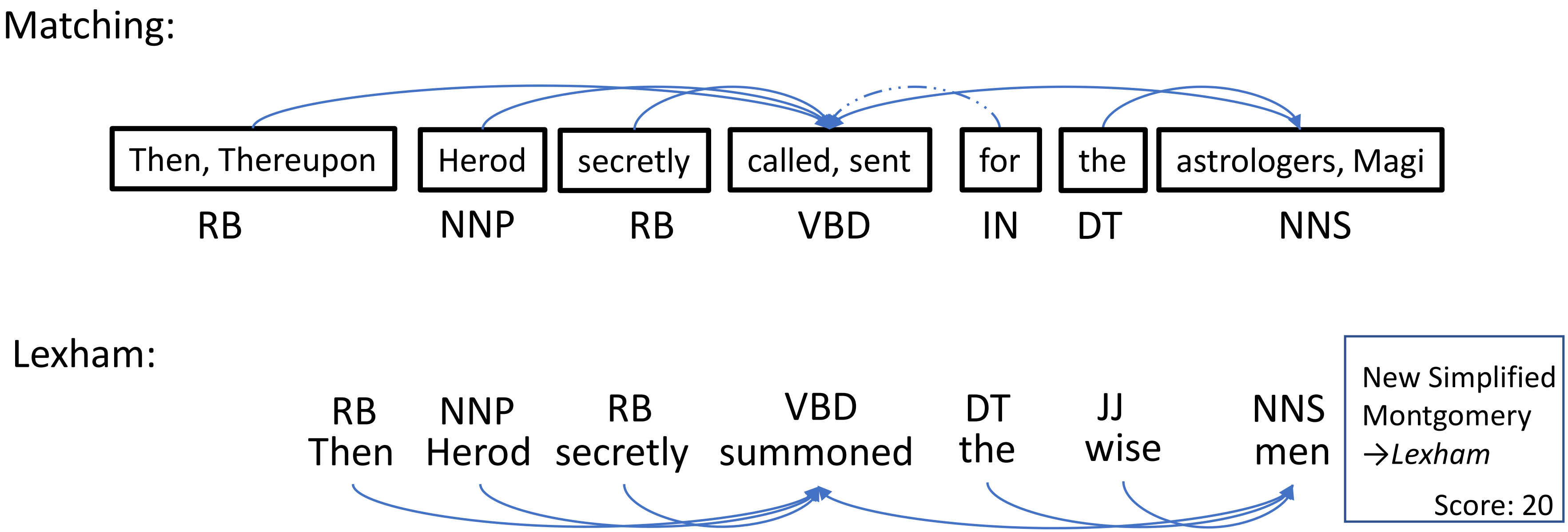
RB	NNP	RB	VBD	DT	JJ	NNS
Then	Herod	secretly	summoned	the	wise	men

New Simplified Montgomery
→*Lexham*

Score: 20

Creating a candidate matching

- Repeat for each document



Matching:

Then, Thereupon	Herod	secretly	called, sent	for	the	astrologers, Magi
RB	NNP	RB	VBD	IN	DT	NNS

Lexham:

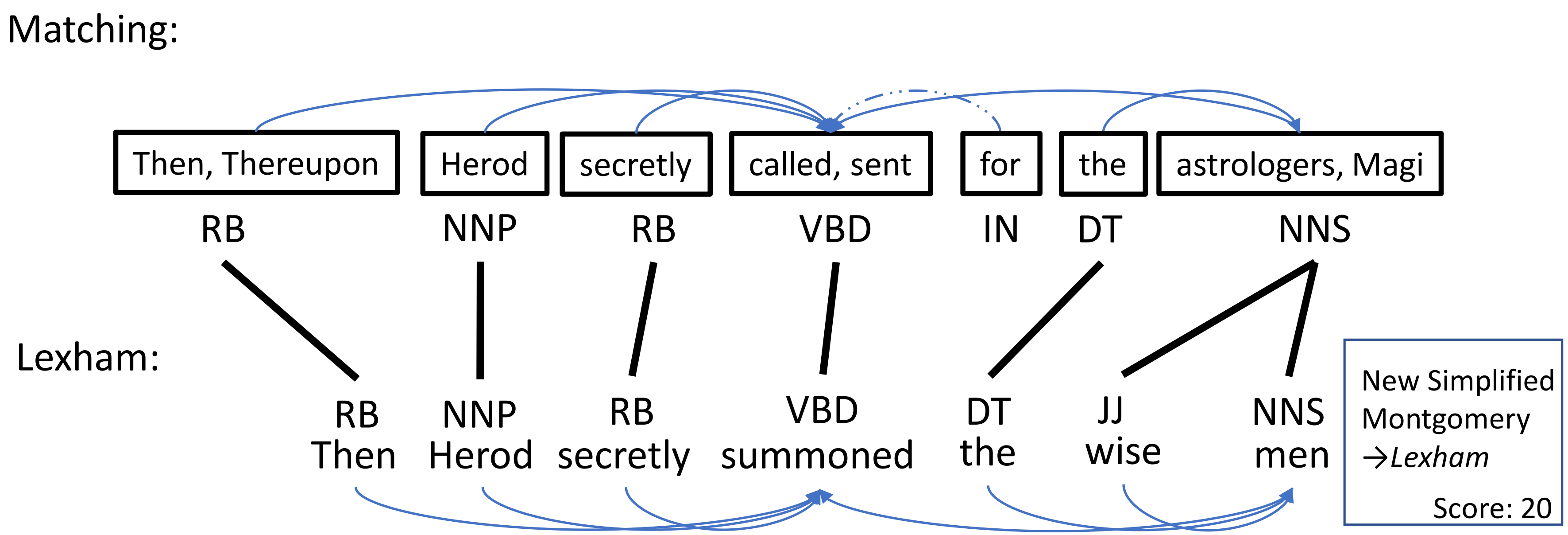
RB	NNP	RB	VBD	DT	JJ	NNS
Then	Herod	secretly	summoned	the	wise	men

New Simplified Montgomery
→*Lexham*

Score: 20

Creating a candidate matching

- Repeat for each document



Matching:

Then, Thereupon | Herod | secretly | called, sent | for | the | astrologers, Magi

RB NNP RB VBD IN DT NNS

Lexham:

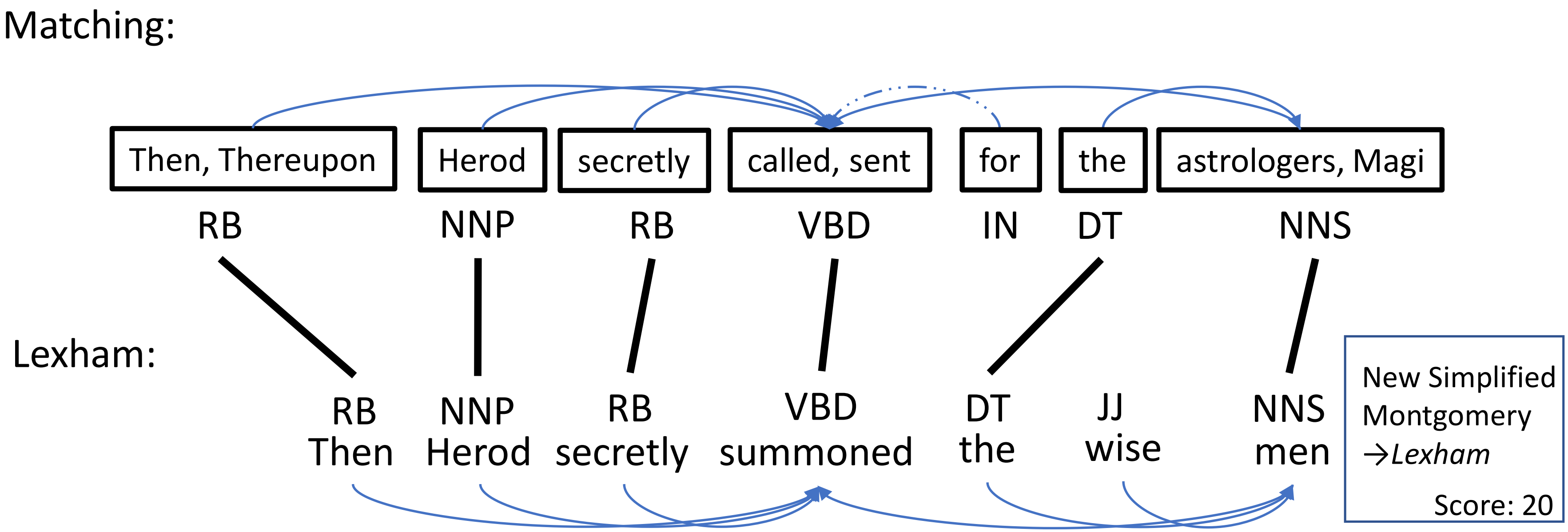
RB Then NNP Herod RB secretly VBD summoned DT the JJ wise NNS men

New Simplified Montgomery →*Lexham*

Score: 20

Creating a candidate matching

- Repeat for each document

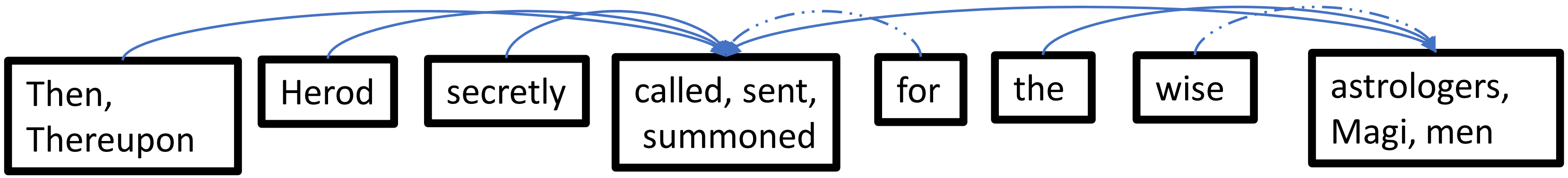


New Simplified Montgomery →*Lexham*

Score: 20

Creating a candidate matching

Matching:



Then, Thereupon	Herod	secretly	called, sent, summoned	for	the	wise	astrologers, Magi, men
RB	NNP	RB	VBD	IN	DT	JJ	NNS

New Simplified
Montgomery
Lexham

Score: 40

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matchings
3. Consensus + Resources

New Simplified
Montgomery
Lexham

Score: 40

Corpus Matchings

Corpus Matchings

- Run for N=10 iterations
    - Rescore edge weights
        - Re-align bitexts with inferred word lists
    - Randomization of document order

Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order

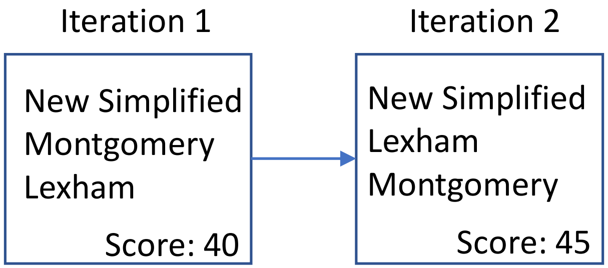
Iteration 1

New Simplified
Montgomery
Lexham

Score: 40

Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order



Iteration 1

New Simplified
Montgomery
Lexham

Score: 40

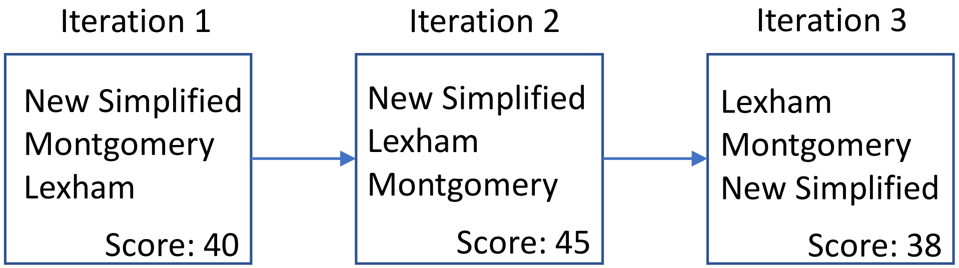
Iteration 2

New Simplified
Lexham
Montgomery

Score: 45

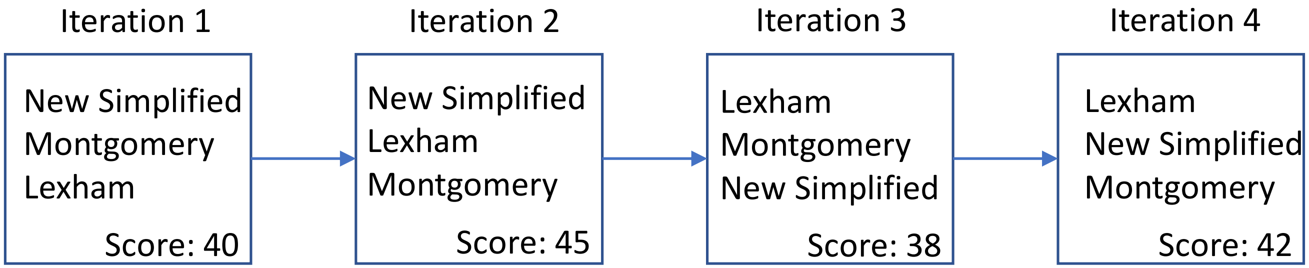
Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order



Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order



Iteration 1

New Simplified
Montgomery
Lexham

Score: 40

Iteration 2

New Simplified
Lexham
Montgomery

Score: 45

Iteration 3

Lexham
Montgomery
New Simplified

Score: 38

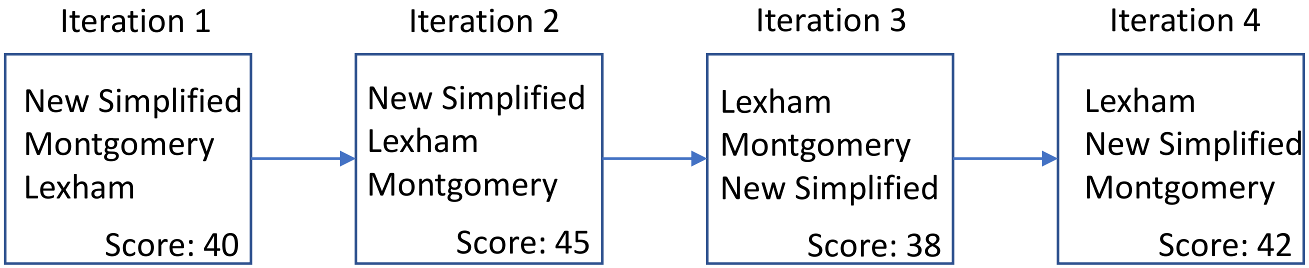
Iteration 4

Lexham
New Simplified
Montgomery

Score: 42

Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order
- For each verse, keep the maximum weight matching



Iteration 1

New Simplified
Montgomery
Lexham

Score: 40

Iteration 2

New Simplified
Lexham
Montgomery

Score: 45

Iteration 3

Lexham
Montgomery
New Simplified

Score: 38

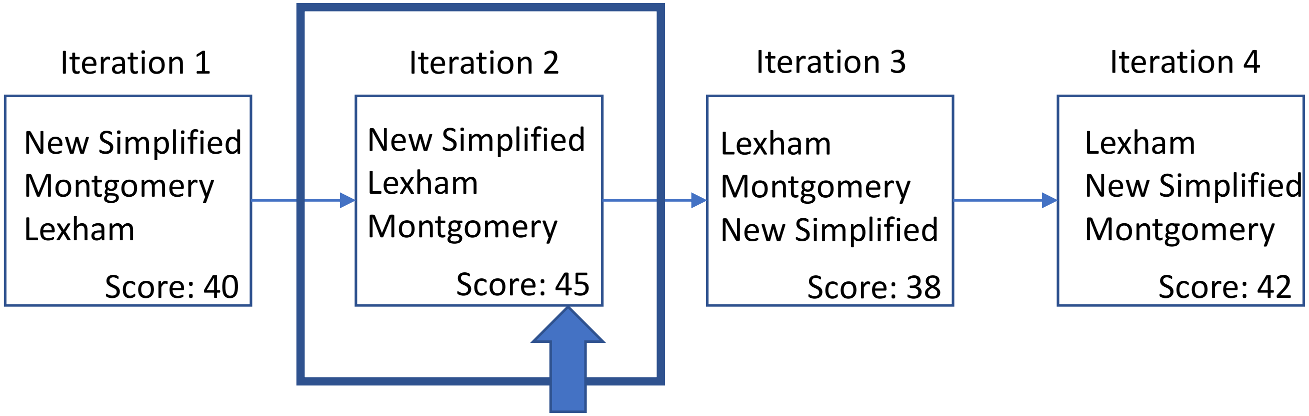
Iteration 4

Lexham
New Simplified
Montgomery

Score: 42

Corpus Matchings

- Run for N=10 iterations
  - Rescore edge weights
    - Re-align bitexts with inferred word lists
  - Randomization of document order
- For each verse, keep the maximum weight matching



Iteration 1

New Simplified
Montgomery
Lexham

Score: 40

Iteration 2

New Simplified
Lexham
Montgomery

Score: 45

Iteration 3

Lexham
Montgomery
New Simplified

Score: 38

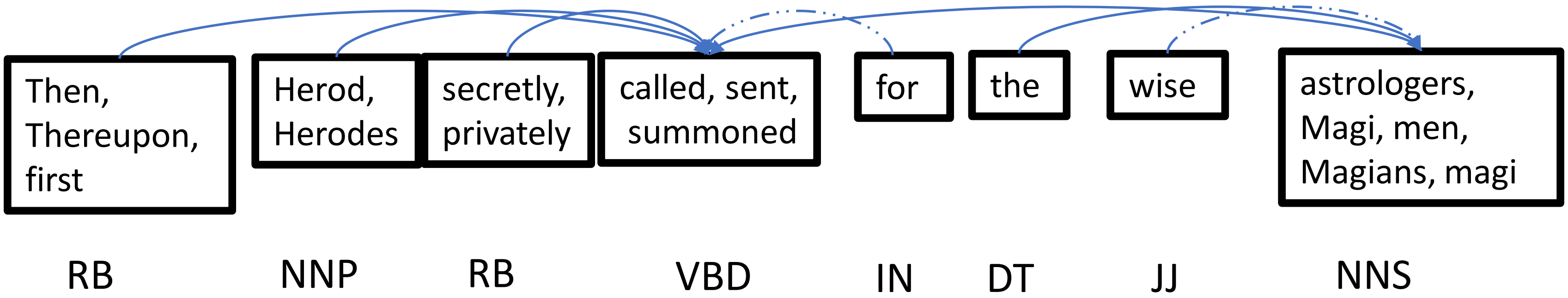
Iteration 4

Lexham
New Simplified
Montgomery

Score: 42

Corpus Matchings

Best matching:



Then, Thereupon, first

Herod, Herodes

secretly, privately

called, sent, summoned

for

the

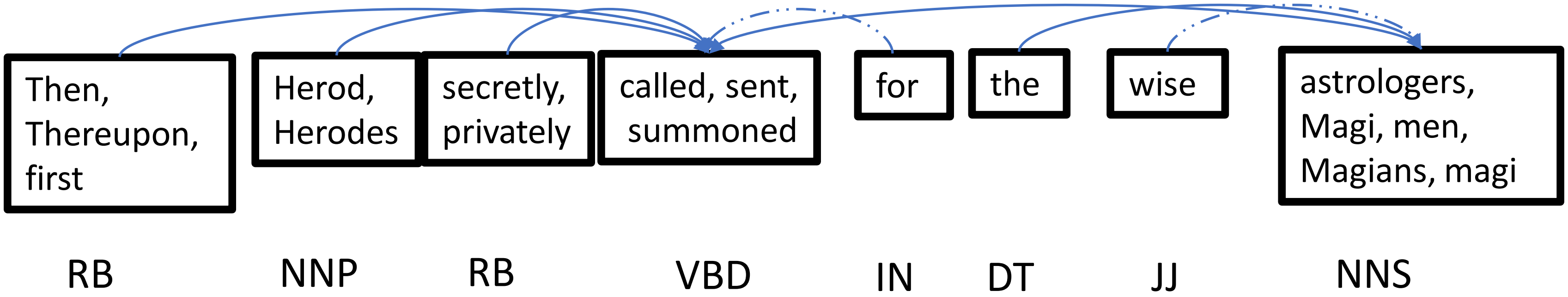
wise

astrologers, Magi, men, Magians, magi

RB NNP RB VBD IN DT JJ NNS

Corpus Matchings

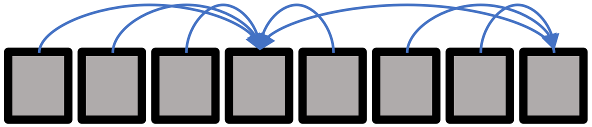
Best matching:



- A Bible in a low-resource language can align or project to a single matching instead of several different versions

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources



Consensus + Resources

- Another format:

worldwide	newsimplified	montgomery	etheridge	godword	majority	lexham	common	contemporary
Then	Then	Thereupon	Then	Then	Then	Then	Then	first
Herod	Herod	Herod	Herod	Herodes	Herod	Herod	Herod	Herod
					,			
secretly	secretly	secretly	privately	secretly	secretly	secretly	secretly	secretly
called	called	sent	called	called	called	summoned	called	called
		for					for	
the	the	the	the	the	the	the	the	the
wise				wise	wise	wise		wise
men	astrologers	Magi	Magians	men	men	men	magi	men

Consensus + Resources

- Another format:

worldwide	newsimplified	montgomery	etheridge	godword	majority	lexham	common	contemporary
Then	**Then**	Thereupon	**Then**	**Then**	**Then**	**Then**	**Then**	**first**
Herod	**Herod**	**Herod**	**Herod**	Herodes	**Herod**	**Herod**	**Herod**	**Herod**
					,			
secretly	**secretly**	**secretly**	privately	**secretly**	**secretly**	**secretly**	**secretly**	**secretly**
called	**called**	sent	**called**	**called**	**called**	summoned	**called**	**called**
		for					For	
the	**the**	**the**	**the**	**the**	**the**	**the**	**the**	**the**
wise				**wise**	**wise**	**wise**		**wise**
men	astrologers	Magi	Magians	**men**	**men**	**men**	Magi	**men**

Consensus + Resources

- Another format:

worldwide	newsimplified	montgomery	etheridge	godword	majority	lexham	common	contemporary
Then	**Then**	Thereupon	**Then**	**Then**	**Then**	**Then**	**Then**	**first**
Herod	**Herod**	**Herod**	**Herod**	Herodes	**Herod**	**Herod**	**Herod**	**Herod**
					,			
secretly	**secretly**	**secretly**	privately	**secretly**	**secretly**	**secretly**	**secretly**	**secretly**
called	**called**	sent	**called**	**called**	**called**	summoned	**called**	**called**
		for					For	
the	**the**	**the**	**the**	**the**	**the**	**the**	**the**	**the**
wise				**wise**	**wise**	**wise**		**wise**
men	astrologers	Magi	Magians	**men**	**men**	**men**	Magi	**men**

Word	Tag
Then	RB
Herod	NNP
,	.
secretly	RB
called	VBD
for	IN
the	DT
wise	JJ
men	NNS

Consensus + Resources

- Another format:

worldwide	newsimplified	montgomery	etheridge	godword	majority	lexham	common	contemporary
Then	**Then**	Thereupon	**Then**	**Then**	**Then**	**Then**	**Then**	**first**
Herod	**Herod**	**Herod**	**Herod**	Herodes	**Herod**	**Herod**	**Herod**	**Herod**
					,			
secretly	**secretly**	**secretly**	privately	**secretly**	**secretly**	**secretly**	**secretly**	**secretly**
called	**called**	sent	**called**	**called**	**called**	summoned	**called**	**called**
		for					For	
the	**the**	**the**	**the**	**the**	**the**	**the**	**the**	**the**
wise				**wise**	**wise**	**wise**		**wise**
men	astrologers	Magi	Magians	**men**	**men**	**men**	Magi	**men**

Word	Tag
Then	RB
Herod	NNP
,	.
secretly	RB
called	VBD
for	IN
the	DT
wise	JJ
men	NNS

Consensus + Resources

- Another format:

worldwide	newsimplified	montgomery	etheridge	godword	majority	lexham	common	contemporary
Then	**Then**	Thereupon	**Then**	**Then**	**Then**	**Then**	**Then**	**first**
Herod	**Herod**	**Herod**	**Herod**	Herodes	**Herod**	**Herod**	**Herod**	**Herod**
					,			
secretly	**secretly**	**secretly**	privately	**secretly**	**secretly**	**secretly**	**secretly**	**secretly**
called	**called**	sent	**called**	**called**	**called**	summoned	**called**	**called**
		for					For	
the	**the**	**the**	**the**	**the**	**the**	**the**	**the**	**the**
wise				**wise**	**wise**	**wise**		**wise**
men	astrologers	Magi	Magians	**men**	**men**	**men**	Magi	**men**

Consensus

Word	Tag
Then	RB
Herod	NNP
,	.
secretly	RB
called	VBD
for	IN
the	DT
wise	JJ
men	NNS

Consensus + Resources

Consensus + Resources

Consensus POS and dependency parses reinforce annotations

Consensus + Resources

Consensus POS and dependency parses reinforce annotations

- Softer Consensus Tags

Consensus + Resources

Consensus POS and dependency parses reinforce annotations

- Softer Consensus Tags
- POS:
  - “TIME”: Sg. Noun (1.00) → Sg. Noun (0.94), Pl. Noun (0.05) …
  - “SECRET”: Sg. Noun (0.54), Adj (0.46) → Adj (0.51), Sg. Noun (0.47) …

Consensus + Resources

Consensus POS and dependency parses reinforce annotations

- Softer Consensus Tags
- POS:
  - “TIME”: Sg. Noun (1.00) → Sg. Noun (0.94), Pl. Noun (0.05) …
  - “SECRET”: Sg. Noun (0.54), Adj (0.46) → Adj (0.51), Sg. Noun (0.47) …
- Lexical Heads:
  - “SECRET”: in (0.28), is (0.06), kept (0.04) … → in (0.32), place (0.05), mystery (0.04)…

Consensus + Resources

Consensus POS and dependency parses reinforce annotations

- Softer Consensus Tags
- POS:
    - “TIME”: Sg. Noun (1.00) → Sg. Noun (0.94), Pl. Noun (0.05) …
    - “SECRET”: Sg. Noun (0.54), Adj (0.46) → Adj (0.51), Sg. Noun (0.47) …
- Lexical Heads:
    - “SECRET”: in (0.28), is (0.06), kept (0.04) … → in (0.32), place (0.05), mystery (0.04)…

Paraphrase Discovery

- “SECRET”: secret (0.59), mystery (0.19), private (0.04) …

Consensus + Resources

Consensus + Resources

Richer exploration and analysis of text meaning

Consensus + Resources

Richer exploration and analysis of text meaning

- Translator/author variation:
  - “HYMENAEUS”: Hymenaues (0.82), Hymenius (0.04), Hymeneus (0.04), Humenaios (0.04) …

Consensus + Resources

Richer exploration and analysis of text meaning

- Translator/author variation:
  - “HYMENAEUS”: Hymenaues (0.82), Hymenius (0.04), Hymeneus (0.04), Humenaios (0.04) …
- Historical language usage:
  - “BLAZES”: burns (0.48), burning (0.33), **burneth** (0.10), blazes (0.04)

Consensus + Resources

Richer exploration and analysis of text meaning

- Translator/author variation:
  - "HYMENAEUS": Hymenaues (0.82), Hymenius (0.04), Hymeneus (0.04), Humenaios (0.04) …
- Historical language usage:
  - "BLAZES": burns (0.48), burning (0.33), **burneth** (0.10), blazes (0.04)
- In-domain word choices and disambiguation:
  - "CHALLENGED": said (0.15), opposed (0.13), urged (0.12), tested (0.10), tempted (0.06) …

Consensus + Resources

Richer exploration and analysis of text meaning

- Translator/author variation:
    - “HYMENAEUS”: Hymenaues (0.82), Hymenius (0.04), Hymeneus (0.04), Humenaios (0.04) …
- Historical language usage:
    - “BLAZES”: burns (0.48), burning (0.33), **burneth** (0.10), blazes (0.04)
- In-domain word choices and disambiguation:
    - “CHALLENGED”: said (0.15), opposed (0.13), urged (0.12), tested (0.10), tempted (0.06) …

Can shed insight into textual or theological debates

Deriving a Consensus

Deriving a Consensus

1. Creating a candidate matching

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources

What’s next?

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources

What’s next?

Up to you! Code/resources/slides available at

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources

What’s next?

Up to you! Code/resources/slides available at

www.github.com/pitrack/monolign

Deriving a Consensus

1. Creating a candidate matching
2. Corpus Matching
3. Consensus + Resources

What’s next?

Up to you! Code/resources/slides available at

www.github.com/pitrack/monolign

Thank You!