From Aari to Zulu:

Massively Multilingual Creation of Language Tools Using Interlinear Glossed Text

Ryan Georgi
IBM Research, San Jose, CA
June 20th, 2016

The Problem

The Problem

(In Three Words)

The Problem

(In Three Words)

Resource Poor Languages

Resource Poor Languages

- Over 7,000 languages on the planet (Lewis, 2016)

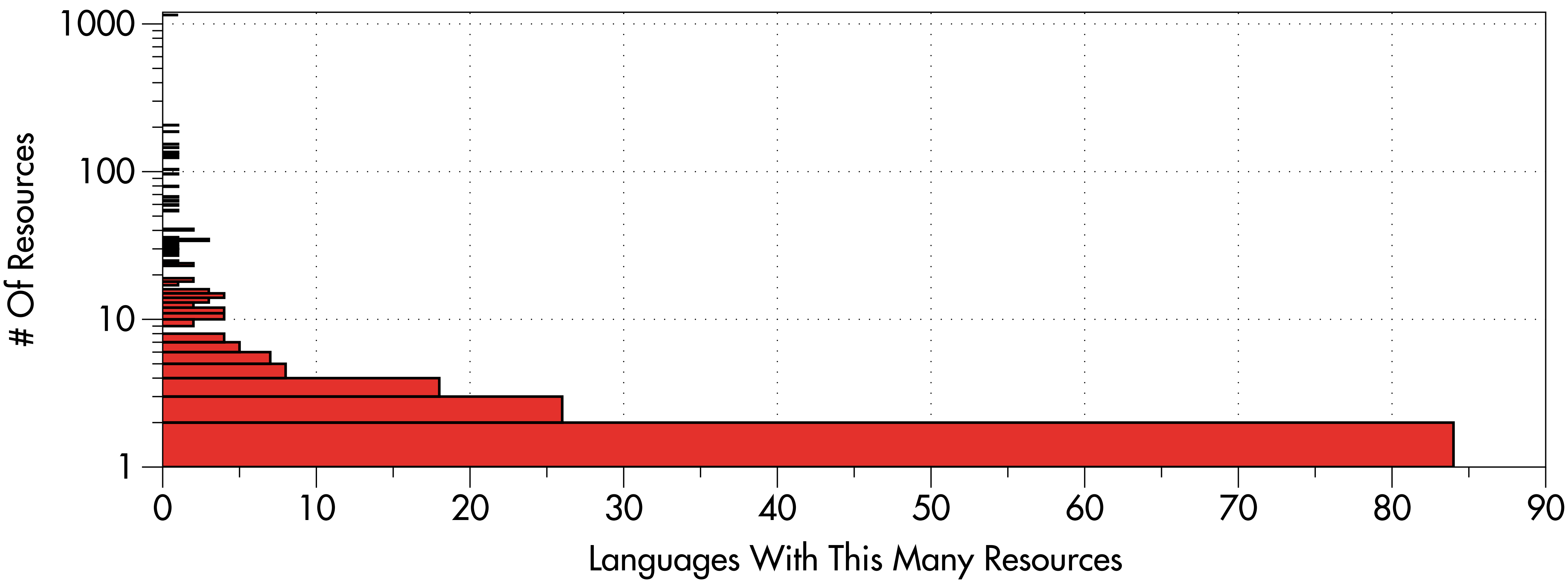
Resource Poor Languages

- Over 7,000 languages on the planet (Lewis, 2016)
- Distribution of language resources very skewed

Resource Poor Languages

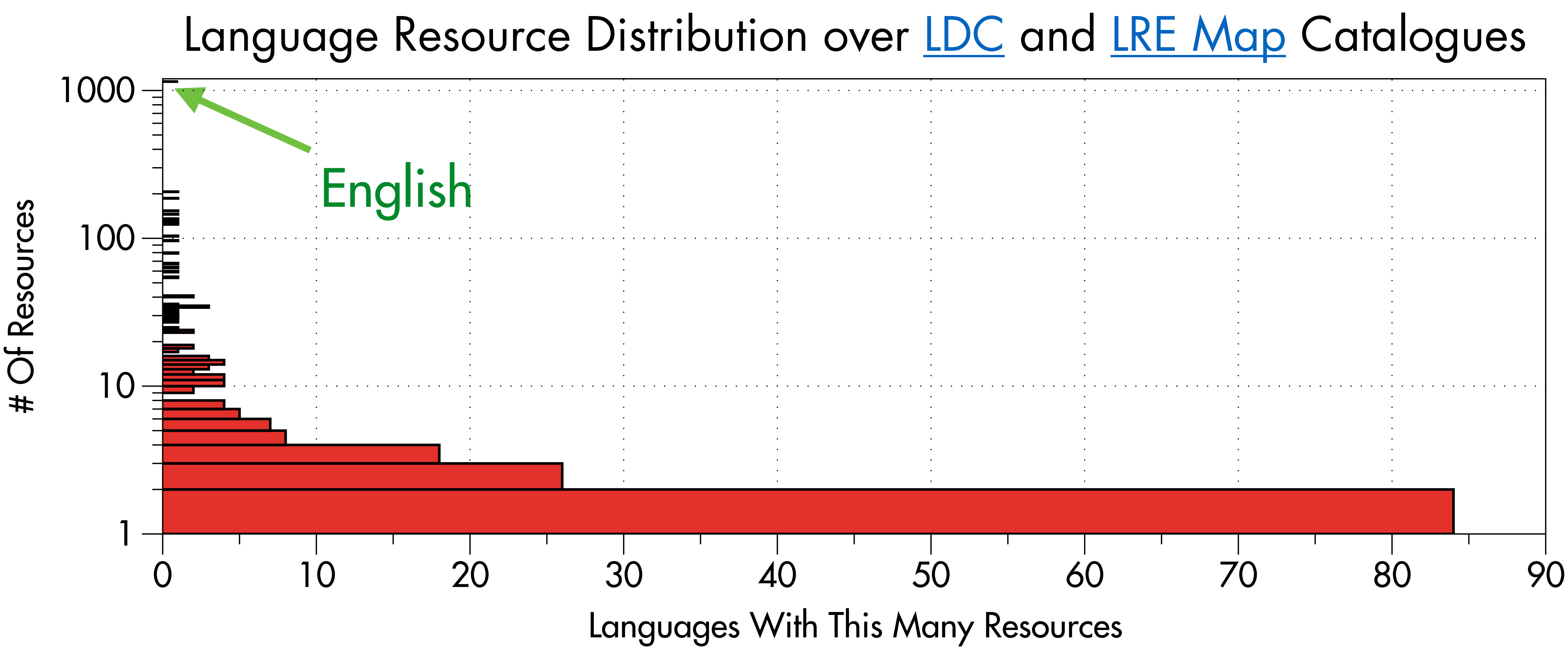
- Over 7,000 languages on the planet (Lewis, 2016)
- Distribution of language resources very skewed

Language Resource Distribution over LDC and LRE Map Catalogues



Resource Poor Languages

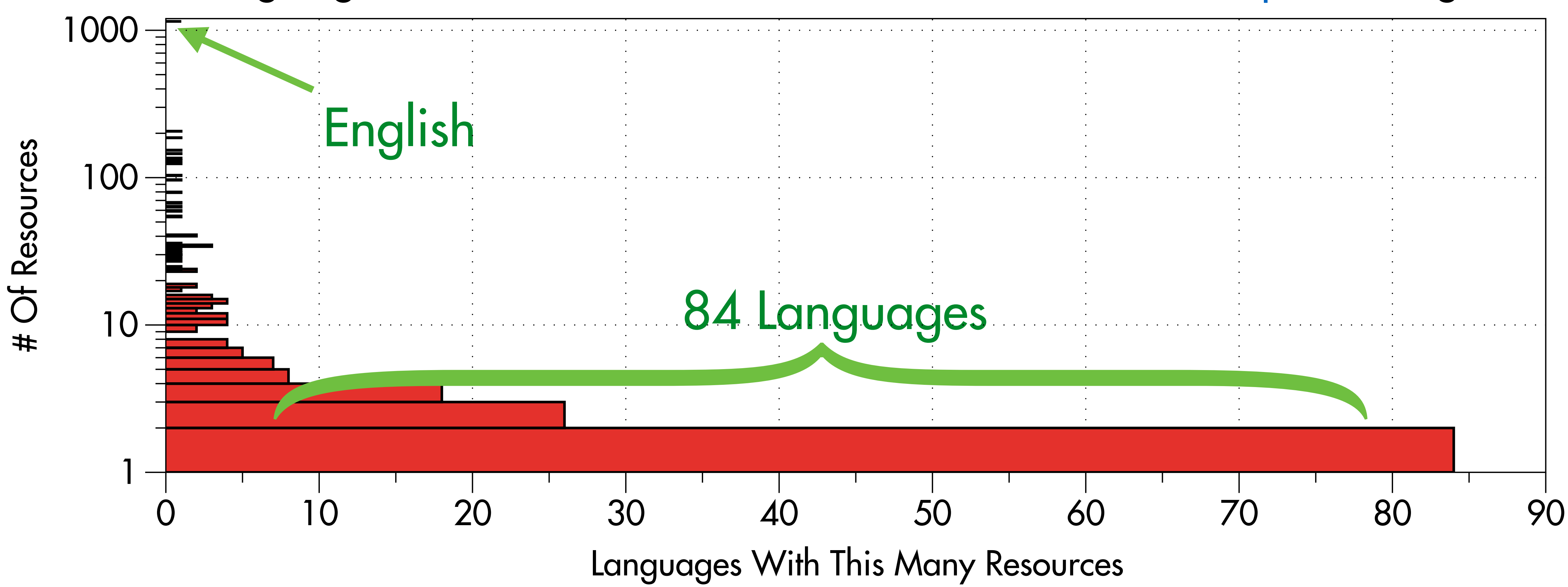
- Over 7,000 languages on the planet ([Lewis, 2016])
- Distribution of language resources very skewed



Resource Poor Languages

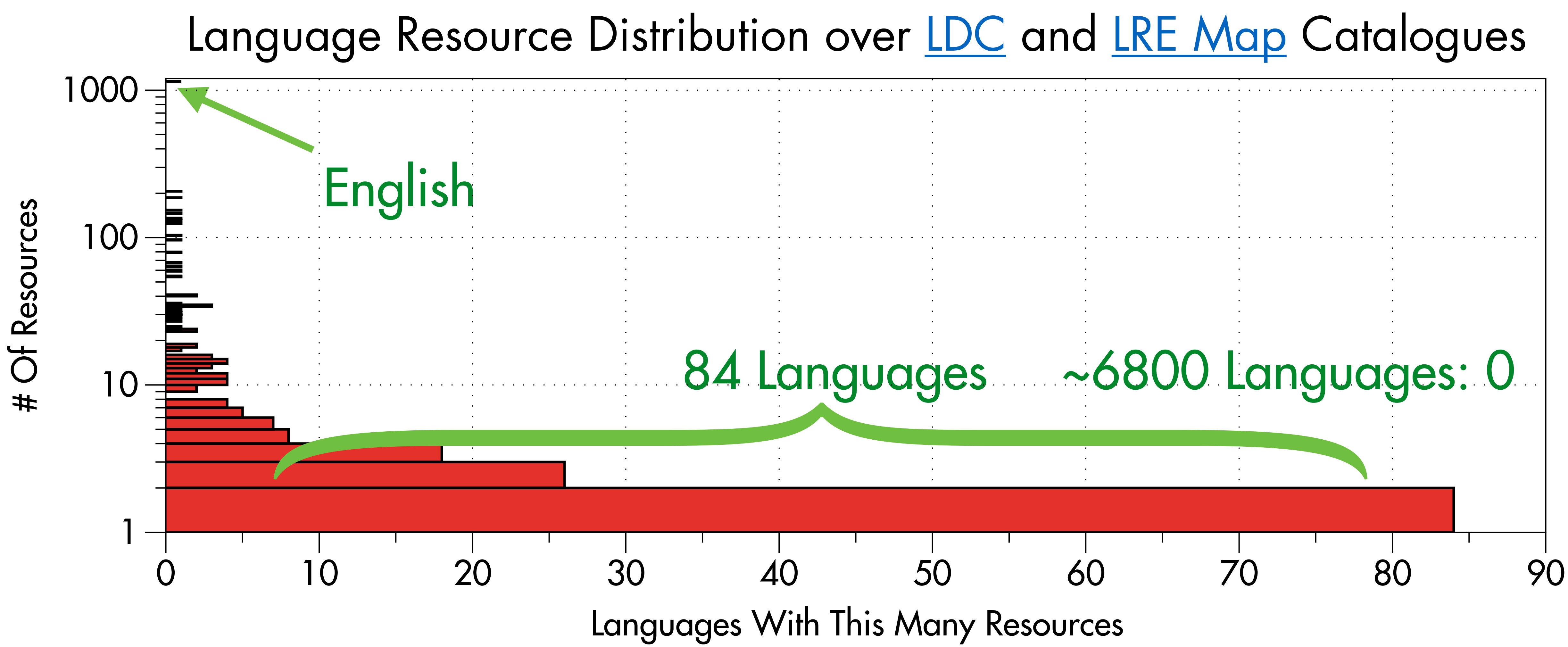
- Over 7,000 languages on the planet (Lewis, 2016)
- Distribution of language resources very skewed

Language Resource Distribution over LDC and LRE Map Catalogues



Resource Poor Languages

- Over 7,000 languages on the planet ([Lewis, 2016])
- Distribution of language resources very skewed



Resource Poor Languages

Resource Poor Languages

- How to approach languages without a large corpus of annotated data?

Resource Poor Languages

- How to approach languages without a large corpus of annotated data?

- Can this approach be **generalizable?**

Outline

- Previous Work
- Methodology
- Tasks
- Conclusion

Previous Work

Previous Work

- Projection from parallel data
  - (Yarowsky & Ngai, 2001; Hwa et. al, 2005)

Previous Work

- Projection from parallel data
  - (Yarowsky & Ngai, 2001; Hwa et. al, 2005)
- Unsupervised, semi-supervised induction
  - Word Clustering (Clark, 2003); Prototypes (Haghighi & Klein, 2006)

Previous Work

- Projection from parallel data
  - (Yarowsky & Ngai, 2001; Hwa et. al, 2005)

- Unsupervised, semi-supervised induction
  - Word Clustering (Clark, 2003); Prototypes (Haghighi & Klein, 2006)

- Delexicalized transfer parsing
  - (Zeman, 2008; McDonald et. al 2011, 2013)

Previous Work

- Projection from parallel data
  - (Yarowsky & Ngai, 2001; Hwa et. al, 2005)
- Unsupervised, semi-supervised induction
  - Word Clustering (Clark, 2003); Prototypes (Haghighi & Klein, 2006)
- Delexicalized transfer parsing
  - (Zeman, 2008; McDonald et. al 2011, 2013)
- Leveraging typological similarities
  - (Hana et. al, 2004; Feldman et. al 2006)

Why a New Approach?

Why a New Approach?

- **Large** quantities of data required for:
  - Statistical alignment
  - Unsupervised/Semi-supervised induction

Why a New Approach?

- **Large** quantities of data required for:
  - Statistical alignment
  - Unsupervised/Semi-supervised induction
- **POS tags** required for transfer parsing approach

Why a New Approach?

- **Large** quantities of data required for:
  - Statistical alignment
  - Unsupervised/Semi-supervised induction
- **POS tags** required for transfer parsing approach
- **Language knowledge** needed for similar language
  - Typological similarity != Genetic Similarity (Georgi et. al 2010)

What is Interlinear Glossed Text?

Interlinear Glossed Text (IGT)

are possible but less common from orders that are not possible at all. Furthermore, languages occasionally exhibit more complex ordering constraints that are not easily represented in such formulae. For example, in Aari (Hayward 1990), an Omotic language spoken in Ethiopia, demonstratives more commonly follow the noun, as in (177a), but they only precede the noun if the noun is followed by a numeral, as in (177b).

(177) b. keené ʔaksí dónq-ine-m
DEM.PLUR dog five-DEF-ACC
'these five dogs'

Aari [aiw] — (Dryer, 2007)

Interlinear Glossed Text (IGT)

are possible but less common from orders that are not possible at all. Furthermore, languages occasionally exhibit more complex ordering constraints that are not easily represented in such formulae. For example, in Aari (Hayward 1990), an Omotic language spoken in Ethiopia, demonstratives more commonly follow the noun, as in (177a), but they only precede the noun if the noun is followed by a numeral, as in (177b).

(177)	b.	keené	?aksí	dónq-ine-m
		DEM.PLUR	dog	five-DEF-ACC
		'these five dogs'		

Aari [aiw] — (Dryer, 2007)

- The **O**nline **D**atabase of **I**nterlinear **T**ext (**ODIN**) *(Lewis & Xia, 2010)*
  - 158,007 IGT instances
  - 1,496 languages
  - 2,027 documents

Why Use IGT?

```
keene     ʔaksí  dónq-ine-m
DEM.PLUR dog    five-DEF-ACC
‘these five dogs’
```

Why Use IGT?

```
keené      ʔaksí  dóŋq-ine-m
DEM.PLUR   dog    five-DEF-ACC
‘these five dogs’
```

- Gloss line contains **grams**

Why Use IGT?

```
keené     ʔaksí dóng-ine-m
DEM.PLUR dog   five-DEF-ACC

‘these five dogs’
```

- Gloss line contains **grams**
- **Morphemes** (when present) often delineated

Why Use IGT?

```
keené     ʔaksí   dónq-ine-m
DEM.PLUR  dog     five-DEF-ACC
‘these five dogs’
```

- Gloss line contains **grams**
- **Morphemes** (when present) often delineated
- Translation and gloss often have **matching tokens**

Why Use IGT?

```
keené      ʔaksí   dónq-ine-m
DEM.PLUR  dog     five-DEF-ACC
‘these five dogs’
```

- Gloss line contains **grams**
- **Morphemes** (when present) often delineated
- Translation and gloss often have **matching tokens**
  - Can be used to align translation with language line

Why Use IGT?

```
PRON       NOUN   NUM
keené      ʔaksí  dónq-ine-m

DEM.PLUR dog      five-DEF-ACC

‘these five dogs’
  PRON  NUM  NOUN
```

- Gloss line contains **grams**
- **Morphemes** (when present) often delineated
- Translation and gloss often have **matching tokens**
  - Can be used to align translation with language line
  - …and “project” information

Outline

- Previous Work
- Methodology
- Tasks
- Conclusion

Main Contributions

- I examine using IGT for three tasks:

Main Contributions

- I examine using IGT for three tasks:
  - **Word Alignment**

Main Contributions

- I examine using IGT for three tasks:
  - **Word Alignment**
  - **Part-of-Speech Tagging**

Main Contributions

- I examine using IGT for three tasks:
  - **Word Alignment**
  - **Part-of-Speech Tagging**
  - **Dependency Parsing**

Main Contributions

Word Alignment

Main Contributions
Word Alignment

- **Heuristic** alignment
  - High precision word alignments with few instances

Main Contributions

Word Alignment

- **Heuristic** alignment
  - High precision word alignments with few instances
- **Statistical** approaches that leverage IGT format
  - Utilize massively multilingual IGT database
  - Demonstrate use of large quantities of IGT data from unrelated languages can improve alignment for resource-poor languages

Main Contributions

Part-of-Speech Tagging

Main Contributions
Part-of-Speech Tagging

- Projection-Based tagging suffers from:
  - Poor word alignments
  - Non-corresponding Projections

Main Contributions

Part-of-Speech Tagging

- Projection-Based tagging suffers from:
  - Poor word alignments
  - Non-corresponding Projections
- Introduce **classification-based** approach
  - Outperforms projection

Main Contributions

Dependency Parsing

- Projection-based parsers compound errors:
  - Word Alignment
  - POS Tagging
  - Non-Correspondance

Main Contributions
Dependency Parsing

- Projection-based parsers compound errors:
  - Word Alignment
  - POS Tagging
  - Non-Correspondance
- Analyze **divergence** to improve parses

Evaluation

Task	Evaluation Settings
Word Alignment	IGT

Evaluation

Task	Evaluation Settings
Word Alignment	IGT
POS Tagging	IGT
	Monolingual
Dependency Parsing	IGT
	Monolingual

Data Overview

Resource Type	ODIN
IGT	✓
POS Tags	
Dependency Structures	
Word Alignment	
# Of Sentences	151,633
# Of Languages	1,487

Data Overview

Resource Type	ODIN	XL-IGT
IGT	✓	✓
POS Tags		
Dependency Structures		✓
Word Alignment		✓
# Of Sentences	151,633	796
# Of Languages	1,487	7

Data Overview

Resource Type	ODIN	XL-IGT	RG-IGT
IGT	✓	✓	✓
POS Tags			✓ Universal
Dependency Structures		✓	
Word Alignment		✓	✓
# Of Sentences	151,633	796	82
# Of Languages	1,487	7	5

Data Overview

Resource Type	ODIN	XL-IGT	RG-IGT	UD-2.0
IGT	✓	✓	✓	
POS Tags			✓ Universal	✓ Universal
Dependency Structures		✓		✓
Word Alignment		✓	✓	
# Of Sentences	151,633	796	82	85,625
# Of Languages	1,487	7	5	8

Data Overview

Resource Type	ODIN	XL-IGT	RG-IGT	UD-2.0	HUTP
IGT	✓	✓	✓		✓
POS Tags			✓ Universal	✓ Universal	✓ Hindi
Dependency Structures		✓		✓	✓
Word Alignment		✓	✓		
# Of Sentences	151,633	796	82	85,625	147
# Of Languages	1,487	7	5	8	1

Data Overview By Language

Family	Language	ISO	ODIN	XL-IGT	RG-IGT	UD-2.0	HUTP
Afroasiatic	Hausa	`hau`	✓	✓			
Austronesian	Indonesian	`ind`	✓			✓	
	Malagasy	`mlg`	✓	✓			
Indo-European	Bulgarian	`bul`	✓		✓		
	French	`fra`	✓		✓	✓	
	Gaelic	`gla`	✓	✓			
	German	`deu`	✓	✓	✓	✓	
	Hindi	`hin`	✓				✓
	Italian	`ita`	✓		✓	✓	
	Spanish	`spa`	✓		✓	✓	
	Swedish	`swe`	✓			✓	
	Welsh	`cym`	✓	✓			
Koreanic	Korean	`kor`	✓	✓		✓	
Uto-Aztecan	Yaqui	`yaq`	✓	✓			

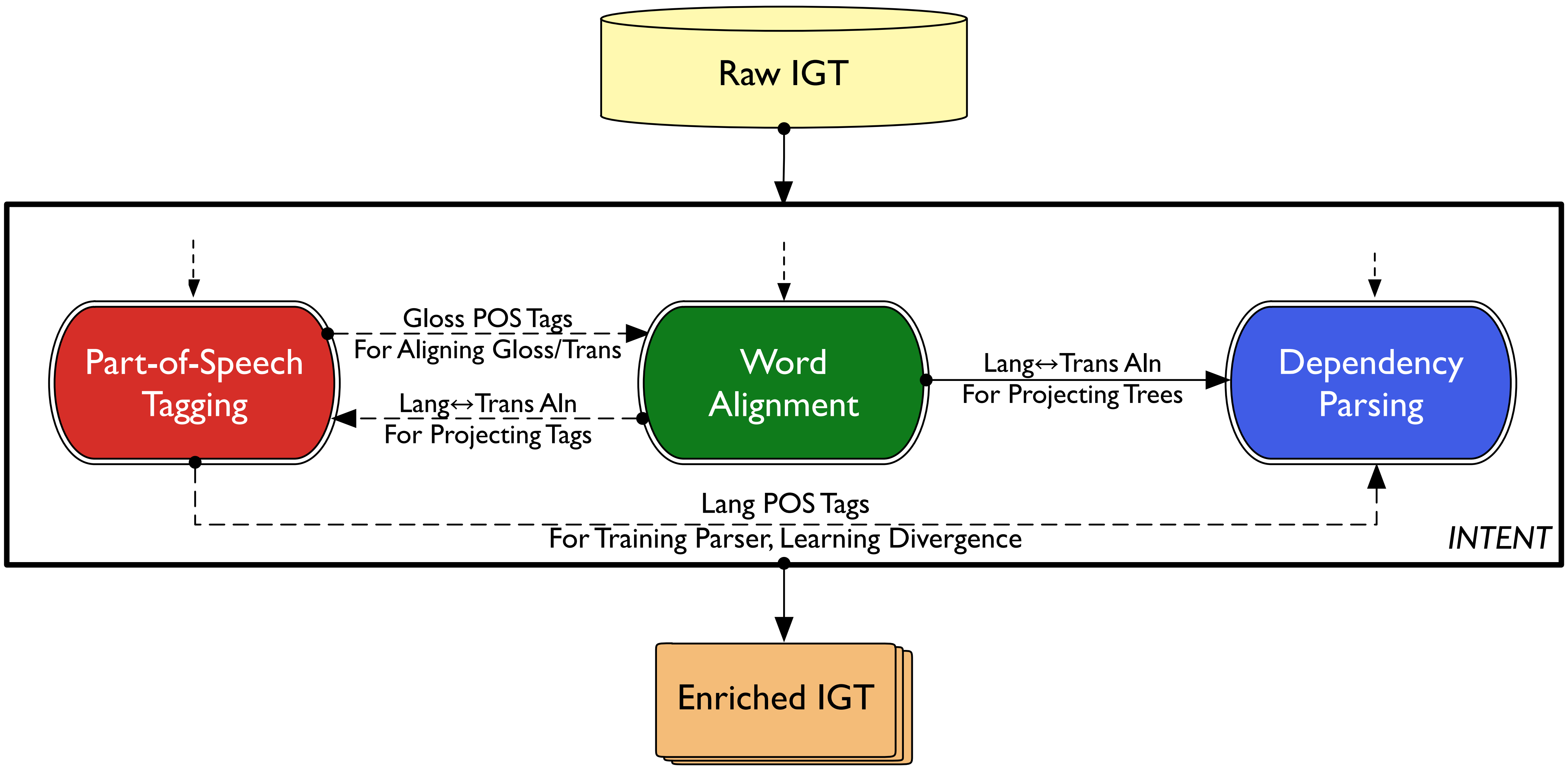
Tasks

Word Alignment

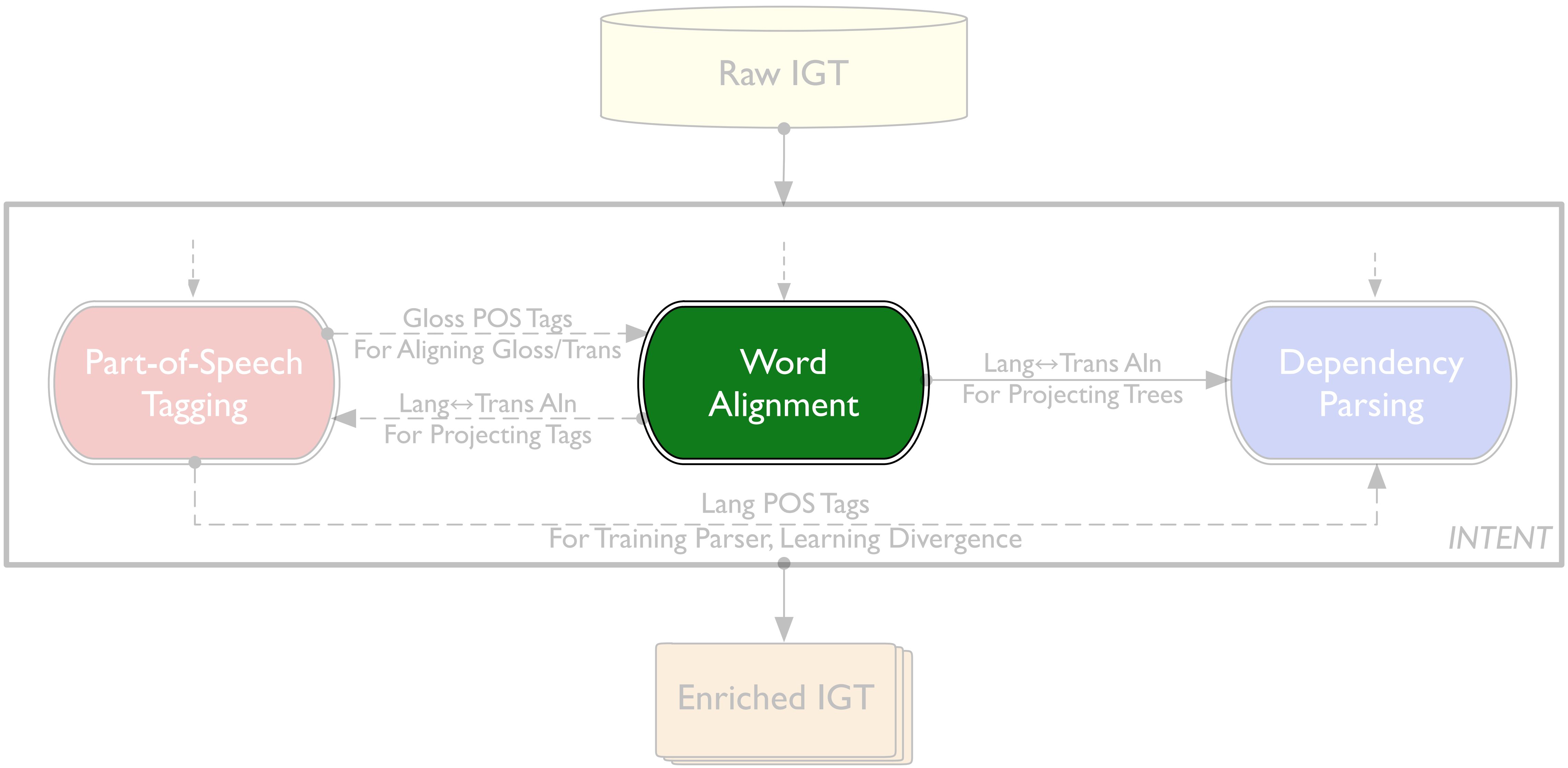
Part-of-Speech Tagging

Dependency Parsing

The INTENT System



The INTENT System



Word Alignment Approaches

- **Heuristic**-based Approach
- **Statistical**-based Approach

Heuristic Word Alignment

0	da	zo-ro	gɛ-rɛ	wuo-ro	la	haane
3SG	PAST	run-IMPERF	go-IMPERF	collect-IMPERF	FACT	berries

He/she was always running there collecting berries.

Dagaare *[dga] (Beerman and Hellan, 2002):*

Heuristic Word Alignment

0	da	zo-ro	gɛ-rɛ	wuo-ro	la	haane
3SG	PAST	run-IMPERF	go-IMPERF	collect-IMPERF	FACT	berries

He/she was always running there collecting berries.

***Dagaare** [dga] (Beerman and Hellan, 2002):*

Heuristic Word Alignment

```
0    da    zo-ro        gɛ-rɛ       wuo-ro           la     haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT   berries
```

He/she was always running there collecting berries.

***Dagaare** [dga] (Beerman and Hellan, 2002):*

Heuristic Word Alignment

```
0    da    zo-ro        gɛ-rɛ      wuo-ro          la    haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT  berries
```

He/she was always running there collecting ***berries****.*

Dagaare *[dga] (Beerman and Hellan, 2002):*

- String matches

Heuristic Word Alignment

```
0    da    zo-ro       gɛ-rɛ      wuo-ro          la    haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT  berries

 He/she was always running there collecting berries.
```

Dagaare *[dga] (Beerman and Hellan, 2002):*

- String matches
- Stemmed String Matches

Heuristic Word Alignment

```
0    da    zo-ro        gɛ-rɛ       wuo-ro          la    haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT  berries

 He/she was always running there collecting berries.
```

Dagaare *[dga] (Beerman and Hellan, 2002):*

- String matches
- Stemmed String Matches
- Word → Gram matches

Heuristic Word Alignment

```
0    da    zo-ro        gɛ-rɛ      wuo-ro          la    haane
3SG  PAST  run-IMPERF   go-IMPERF  collect-IMPERF  FACT  berries
```

He/she was always running there collecting berries.

***Dagaare** [dga] (Beerman and Hellan, 2002):*

- String matches
- Stemmed String Matches
- Word → Gram matches
- Unmatched Tokens

Statistical Word Alignment

- Two targets for translation line:

Statistical Word Alignment

- Two targets for translation line:
  - **L-T**: Language → Translation Alignment
    - Use L/T sentence pairs from the given language

Statistical Word Alignment

- Two targets for translation line:
  - **L-T**: Language → Translation Alignment
    - Use L/T sentence pairs from the given language
  - **G-T**: Gloss → Translation Alignment

Statistical Word Alignment

- Two targets for translation line:
  - **L-T**: Language → Translation Alignment
    - Use L/T sentence pairs from the given language
  - **G-T**: Gloss → Translation Alignment
    - Gloss line is cross-linguistic “pseudo-language”

Statistical Word Alignment

- Two targets for translation line:
  - **L-T**: Language → Translation Alignment
    - Use L/T sentence pairs from the given language
  - **G-T**: Gloss → Translation Alignment
    - Gloss line is cross-linguistic "pseudo-language"
    - Can use G/T sentence pairs from **ALL** languages

Statistical Word Alignment

- Two targets for translation line:
  - **L-T**: Language → Translation Alignment
    - Use L/T sentence pairs from the given language
  - **G-T**: Gloss → Translation Alignment
    - Gloss line is cross-linguistic "pseudo-language"
    - Can use G/T sentence pairs from **ALL** languages
    - (**G-T+ALL ODIN**)

Combining Statistical & Heuristic Alignment

L-T

G-T

G-T + ALL ODIN

- Add word pairs from heuristic aligner to training data

Combining Statistical & Heuristic Alignment

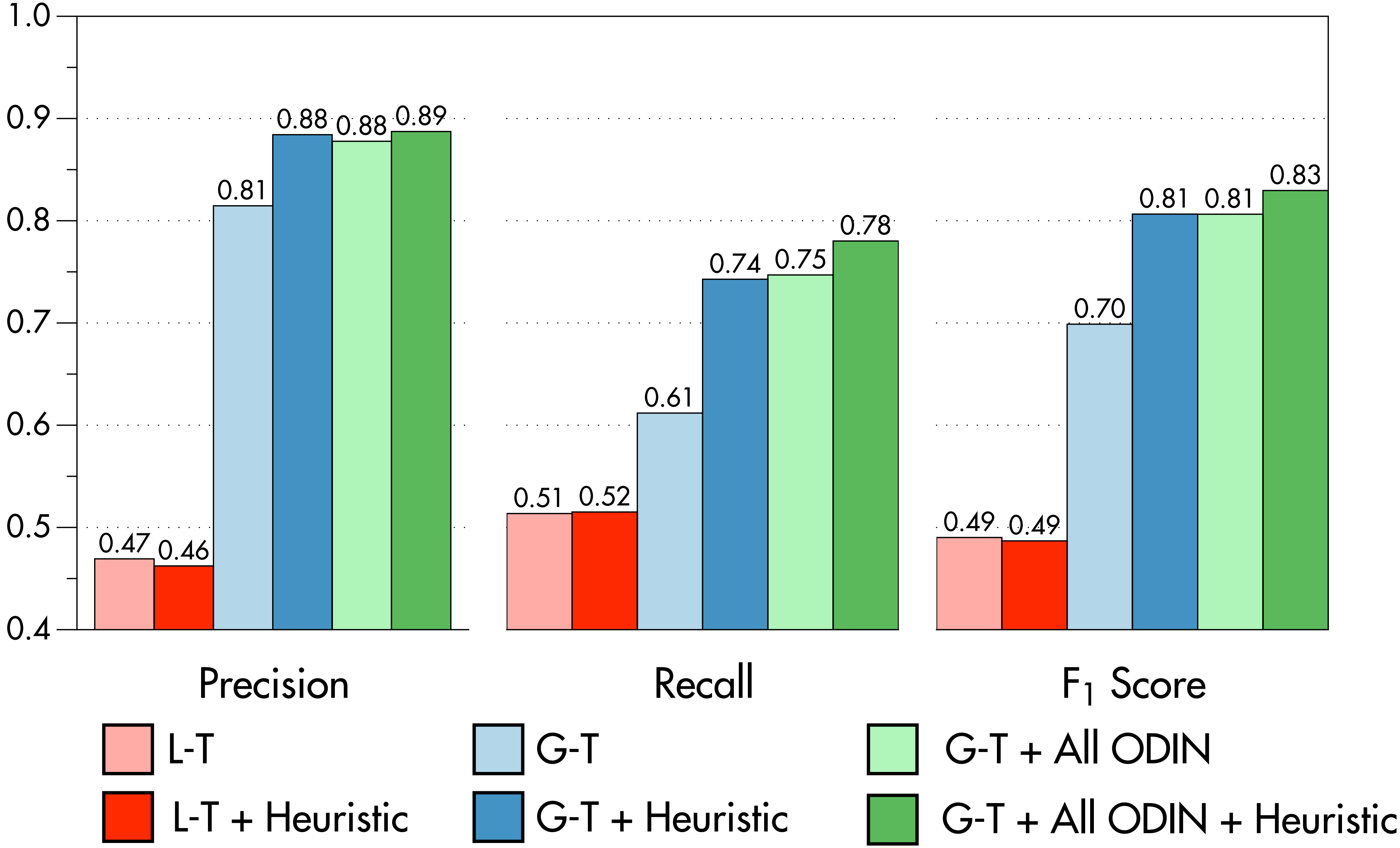
L-T | **L-T + Heuristic**

G-T | **G-T + Heuristic**

G-T + ALL ODIN | **G-T + ALL ODIN + Heuristic**

- Add word pairs from heuristic aligner to training data

Word Alignment



POS Tag Heuristic

```
a. Piarresek egin  du   etchea.
   Peter-ERG make  has  house-ABS

  "Peter built the house."
```

Basque [eus] — (Lafitte, 1962)

POS Tag Heuristic

`Peter-ERG make  has  house-ABS`

"Peter built the house."

POS Tag Heuristic

```
Peter-ERG make  has  house-ABS

"Peter built the house."
```

POS Tag Heuristic

```
Peter-ERG make  has  house-ABS
```

"Peter built the house."

NOUN ***VERB*** ***DET*** ***NOUN***

POS Tag Heuristic

```
NOUN         VERB   VERB  NOUN
Peter-ERG make   has   house-ABS

"Peter built the house."
 NOUN   VERB   DET NOUN
```

POS Tag Heuristic

VERB ***VERB***
make has

built the
VERB ***DET***

POS Tag Heuristic

```
VERB  VERB
make  has

built the
VERB  DET
```

POS Tag Heuristic

`Peter-ERG make  has  house-ABS`

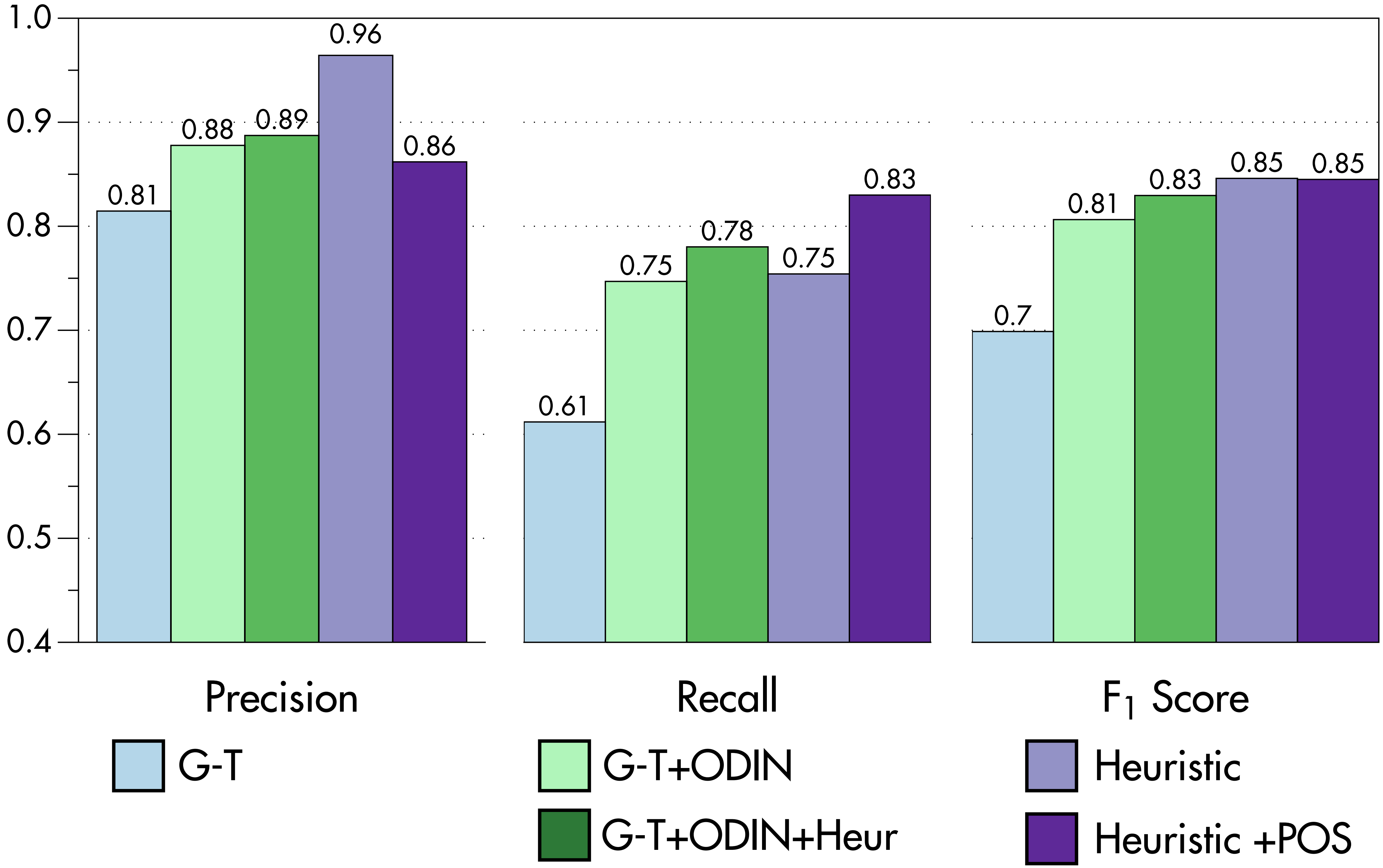
"Peter ***built*** *the house."*

POS Tag Heuristic

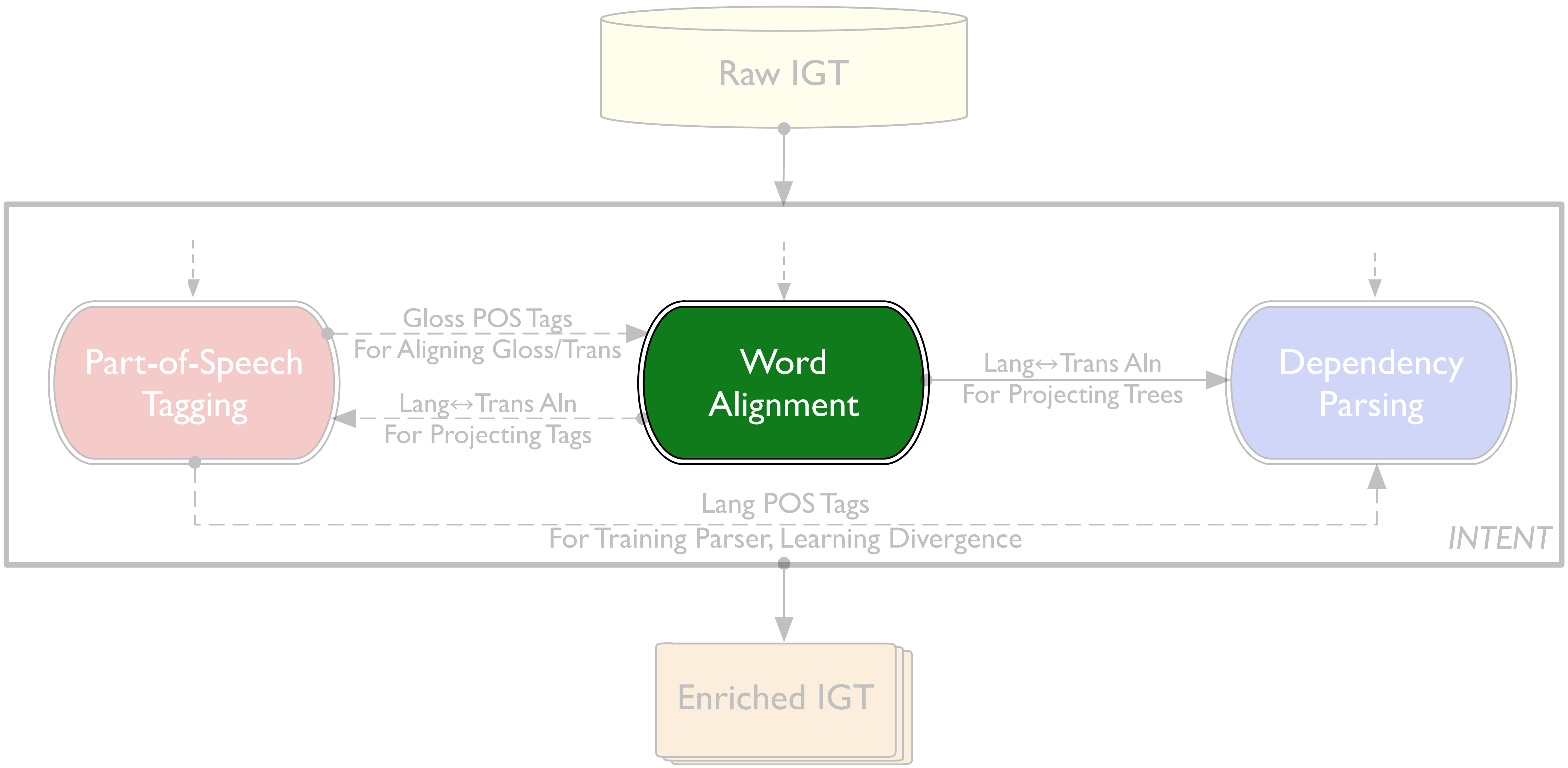
```
a. Piarresek egin  du   etchea.
   Peter-ERG make  has  house-ABS
```

"Peter built the house."

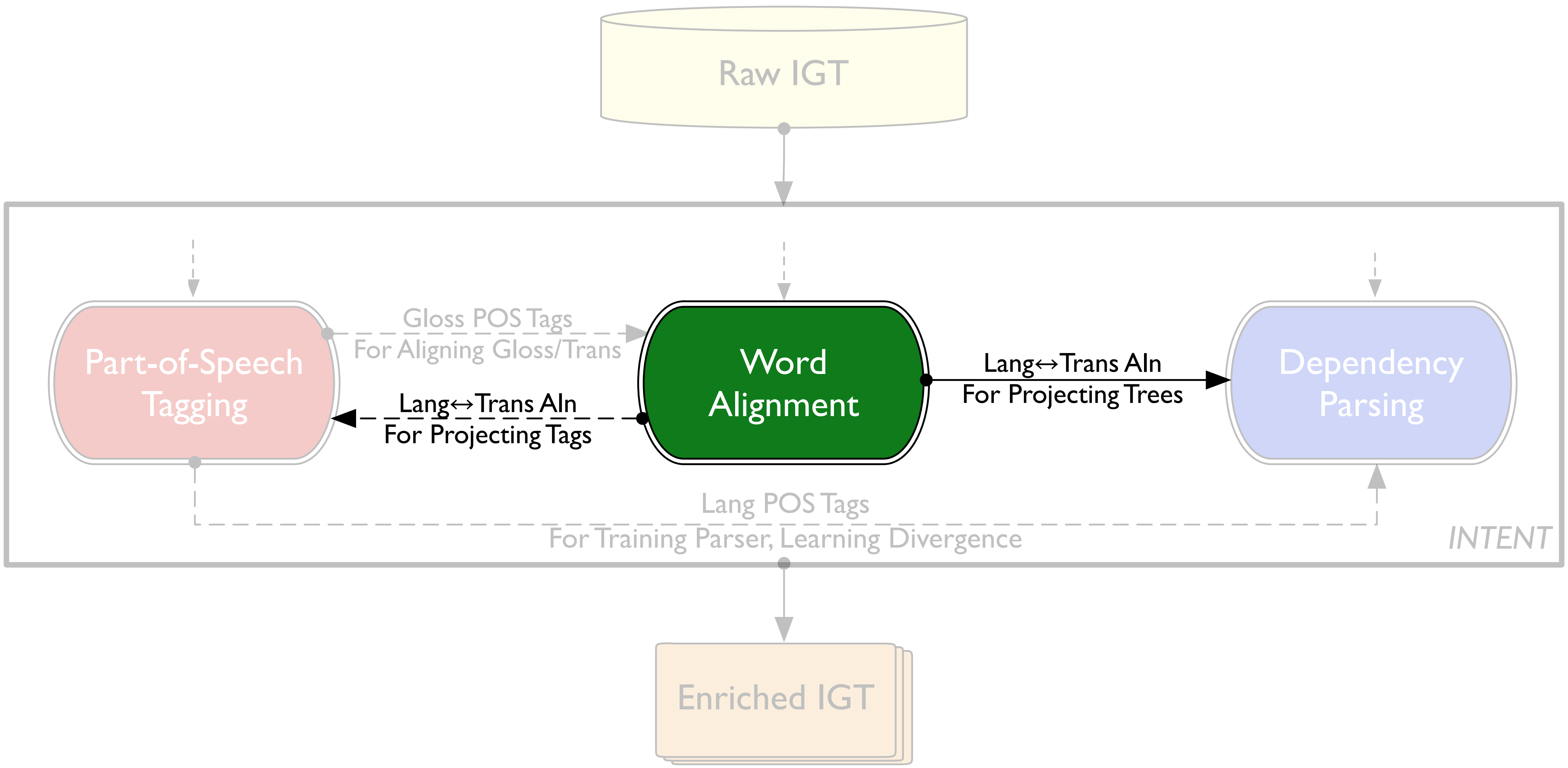
Word Alignment



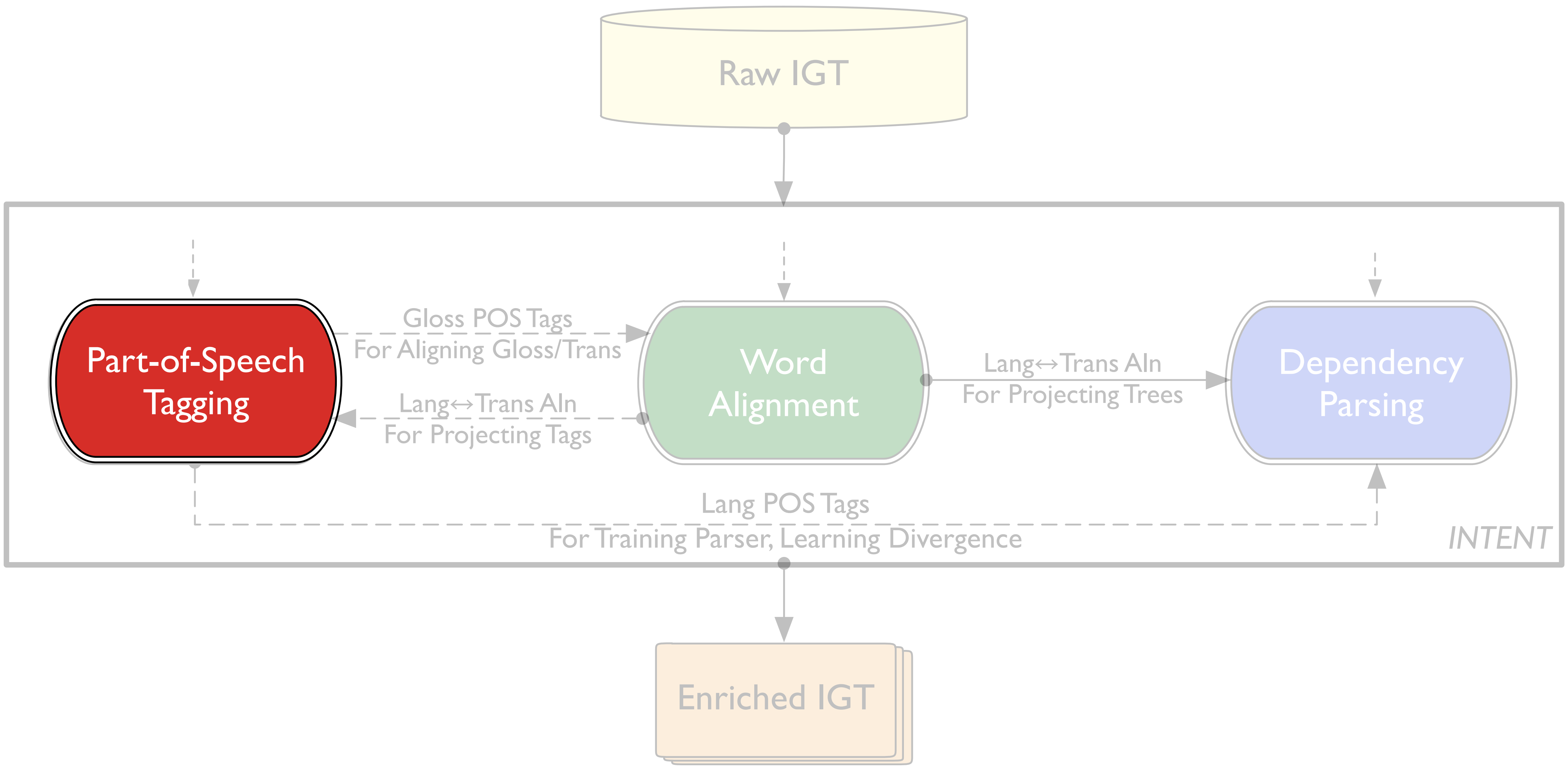
The INTENT System



The INTENT System



The INTENT System



POS Projection

```
0    da    zo-ro        gɛ-rɛ       wuo-ro            la     haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT   berries
```

He/she was always running there collecting berries.

POS Projection

```
0    da    zo-ro        gɛ-rɛ     wuo-ro          la    haane
3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT  berries
```

He/she was always running there collecting berries.
PRON VERB ADV VERB PRON VERB NOUN

- Use English POS tagger

POS Projection

0	da	zo-ro	gɛ-rɛ	wuo-ro	la	haane
3SG	PAST	run-IMPERF	go-IMPERF	collect-IMPERF	FACT	berries

He/she was always running there collecting berries.

PRON VERB ADV VERB PRON VERB NOUN

- Use English POS tagger
- Obtain word alignment

POS Projection

```
PRON  VERB    VERB                          VERB                       NOUN
 0     da    zo-ro        gɛ-rɛ      wuo-ro                la     haane
 3SG  PAST  run-IMPERF  go-IMPERF  collect-IMPERF  FACT   berries

  He/she was always running there collecting berries.
   PRON  VERB   ADV      VERB    PRON     VERB        NOUN
```

- Use English POS tagger
- Obtain word alignment
- Project POS tags to language line.

POS Projection

```
                            ??                           ??
0    da    zo-ro            gɛ-rɛ     wuo-ro             la     haane
3SG  PAST  run-IMPERF       go-IMPERF collect-IMPERF     FACT   berries
```

He/she was always running there collecting berries.

- Use English POS tagger
- Obtain word alignment
- Project POS tags to language line.
- Words that remain unaligned:
  - Tag with "UNK"?
  - Tag with most common tag? (NOUN?)

POS Tagging

- A few unaligned words is fine, but can be worse:

Chintang [ctn] (Bickel et. al, 2007):

```
numphurɨk bhir-ce       mett-ma-ce            par-ch-a
a.place   precipice-ns  do.with/to-INF-3nsP  must-NPST-3s
We have to be sensible  about the Namphuruk cliff.
```

POS Tagging

- A few unaligned words is fine, but can be worse:

Chintang [ctn] (Bickel et. al, 2007):

```
numphurɨk bhir-ce        mett-ma-ce           par-ch-a
a.place   precipice-ns   do.with/to-INF-3nsP  must-NPST-3s
We have to be sensible   about the Namphuruk cliff.
```

- No clear heuristic alignment

POS Tagging

- A few unaligned words is fine, but can be worse:

Chintang [ctn] (Bickel et. al, 2007):

```
numphurɨk bhir-ce        mett-ma-ce              par-ch-a
a.place   precipice-ns   do.with/to-INF-3nsP   must-NPST-3s
We have to be sensible   about the Namphuruk cliff.
```

- No clear heuristic alignment
- …but still plenty of clues in gloss

POS Tagging

- A few unaligned words is fine, but can be worse:

Chintang [ctn] (Bickel et. al, 2007):

```
numphurɨk bhir-ce        mett-ma-ce             par-ch-a
a.place   precipice-ns   do.with/to-INF-3nsP    must-NPST-3s
We have to be sensible   about the Namphuruk cliff.
```

- No clear heuristic alignment
- …but still plenty of clues in gloss

Gloss-Line Feature Extraction

a.place

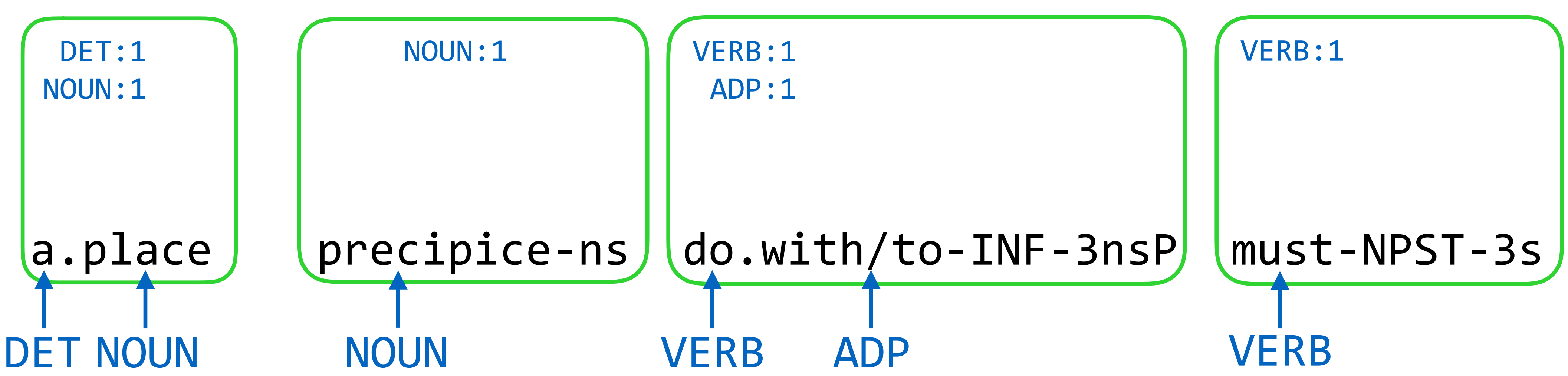
precipice-ns

do.with/to-INF-3nsP

must-NPST-3s

Gloss-Line Feature Extraction

- Most common tag for English words in gloss



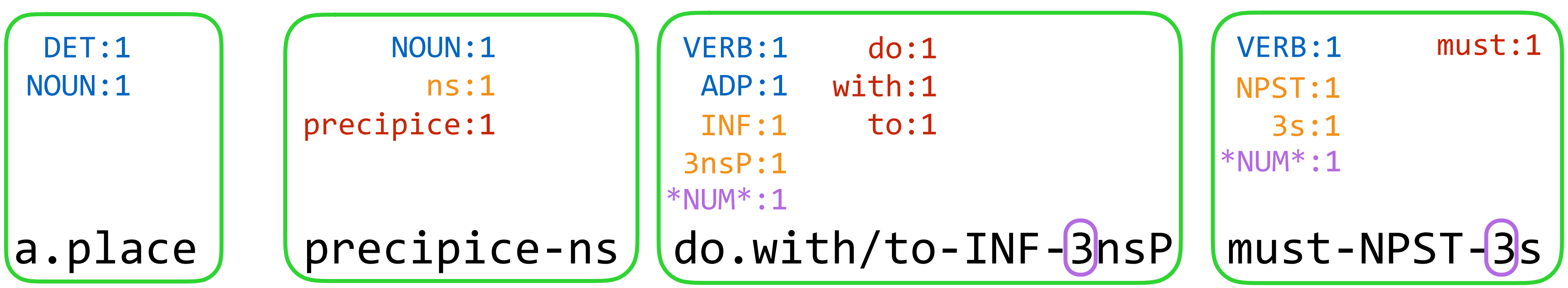
Gloss-Line Feature Extraction

- Most common tag for English words in gloss
- Each "sub-word," including grams

a.place	precipice-ns	do.with/to-INF-3nsP	must-NPST-3s
DET:1 NOUN:1	NOUN:1 ns:1 precipice:1	VERB:1 ADP:1 INF:1 3nsP:1 do:1 with:1 to:1	VERB:1 NPST:1 3s:1 must:1

Gloss-Line Feature Extraction

- Most common tag for English words in gloss
- Each “sub-word,” including grams
- Has a Number



Gloss-Line Feature Extraction

- Most common tag for English words in gloss
- Each "sub-word," including grams
- Has a Number
- …and more

```
DET:1
NOUN:1
a.place
```

```
NOUN:1
ns:1
precipice:1
precipice-ns
```

```
VERB:1      do:1
ADP:1       with:1
INF:1       to:1
3nsP:1
*NUM*:1
do.with/to-INF-3nsP
```

```
VERB:1      must:1
NPST:1
3s:1
*NUM*:1
must-NPST-3s
```

All Features

`subWords`	[.] [-] or [=] delineated tokens
`alignedTag`	Tag for heuristically aligned translation word
`wordHasNumber`	Contains a numeral
`suffix`	last 1,2,3 characters of word
`prefix`	first 1,2,3 characters of word
`numSubwords`	# of subWords
`prevSubwords`	subWords in previous token
`nextSubwords`	subWords in following token
`dictTag`	If subWord is English: most frequent POS tag
`prevDictTag`	dictTag for prev word
`nextDictTag`	dictTag for next word

All Features

`subWords`	[.] [-] or [=] delineated tokens
`alignedTag`	Tag for heuristically aligned translation word
`wordHasNumber`	Contains a numeral
`suffix`	last 1,2,3 characters of word
`prefix`	first 1,2,3 characters of word
`numSubwords`	# of subWords
`prevSubwords`	**subWords in previous token**
`nextSubwords`	**subWords in following token**
`dictTag`	If subWord is English: most frequent POS tag
`prevDictTag`	**dictTag for prev word**
`nextDictTag`	**dictTag for next word**

Obtaining Labeled Training Data

Obtaining Labeled Training Data

- Manual Annotation

Obtaining Labeled Training Data

- Manual Annotation
- Automatic Projection

Obtaining Labeled Training Data

- Manual Annotation
- Automatic Projection

```
U-mfana    u-zo-fund-a            i-ncwadi.
1-1.boy    1.sbj-fut-study-fv     9-9.book
The boy will study the book.
```

Zulu *[zul] – (Buell, 2003)*

Obtaining Labeled Training Data

- Manual Annotation
- Automatic Projection

```
U-mfana    u-zo-fund-a             i-ncwadi.
1-1.boy    1.sbj-fut-study-fv      9-9.book
The boy will study the book.
    NOUN       VERB       NOUN
```

Zulu *[zul] – (Buell, 2003)*

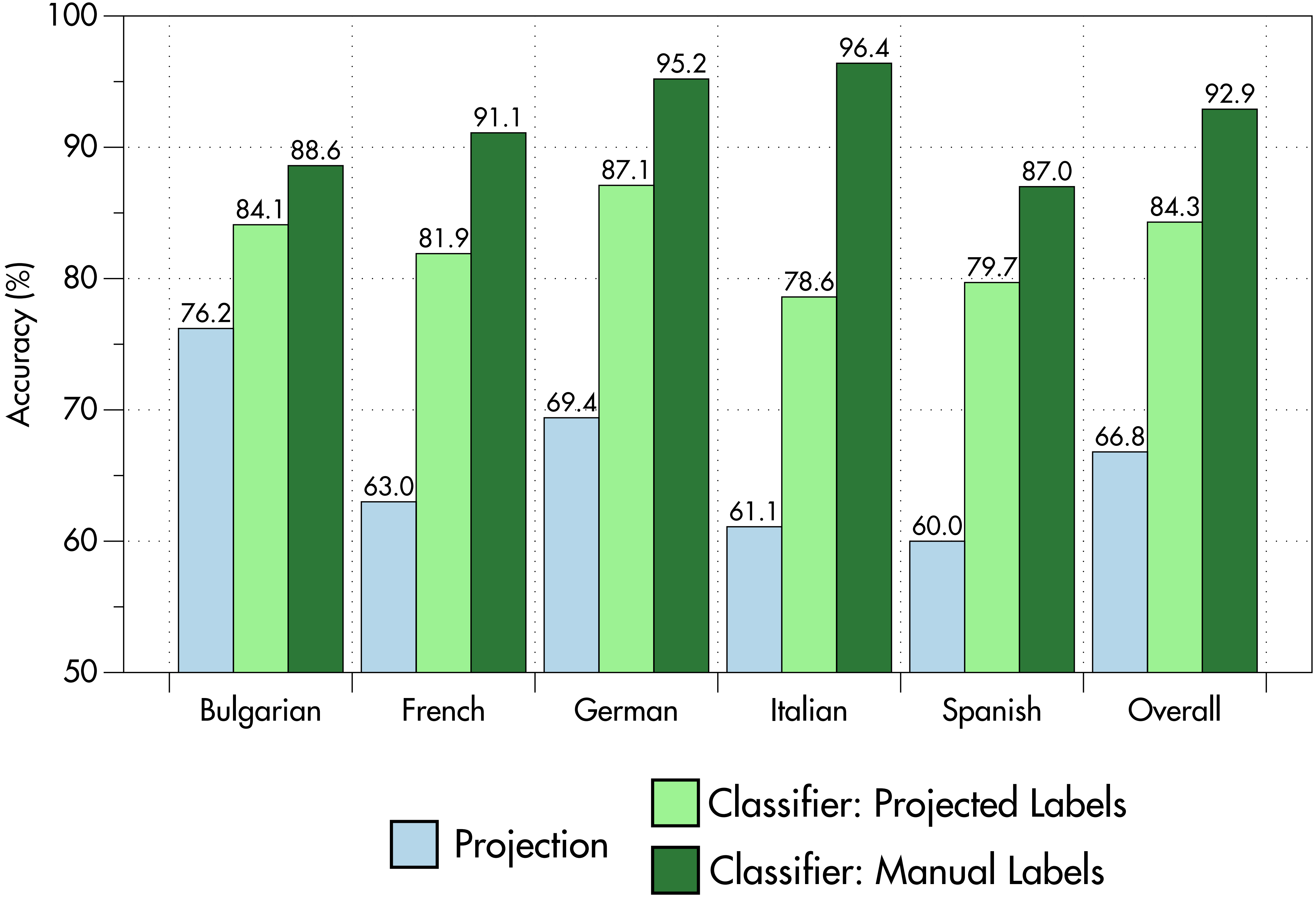
Obtaining Labeled Training Data

- Manual Annotation
- Automatic Projection

NOUN	VERB	NOUN
1-1.boy	1.sbj-fut-study-fv	9-9.book

Zulu *[zul] – (Buell, 2003)*

POS Tagging Results: IGT



POS Tagging

- Now have POS tags on language of interest

POS Tagging

- Now have POS tags on language of interest
- POS Tagging IGT instances interesting, but limited

POS Tagging

- Now have POS tags on language of interest
- POS Tagging IGT instances interesting, but limited
- More general application: novel monolingual data

POS Tagging

- Now have POS tags on language of interest
- POS Tagging IGT instances interesting, but limited
- More general application: novel monolingual data
  - Use POS tags from language line to train monolingual tagger

POS Tagging

- Now have POS tags on language of interest
- POS Tagging IGT instances interesting, but limited
- More general application: novel monolingual data
  - Use POS tags from language line to train monolingual tagger
  - Evaluate w/Universal Dependency Treebank (McDonald et. al, 2013)

Monolingual POS Tagging

- Four settings:

Monolingual POS Tagging

- Four settings:
    - Projection

Monolingual POS Tagging

- Four settings:
  - Projection
    - **All** instances: **with** unaligned words

Monolingual POS Tagging

- Four settings:
  - Projection
    - **All** instances: **with** unaligned words
    - **Filtered** instances: **no** unaligned words

Monolingual POS Tagging

- Four settings:
  - Projection
    - **All** instances: **with** unaligned words
    - **Filtered** instances: **no** unaligned words
  - Classification:

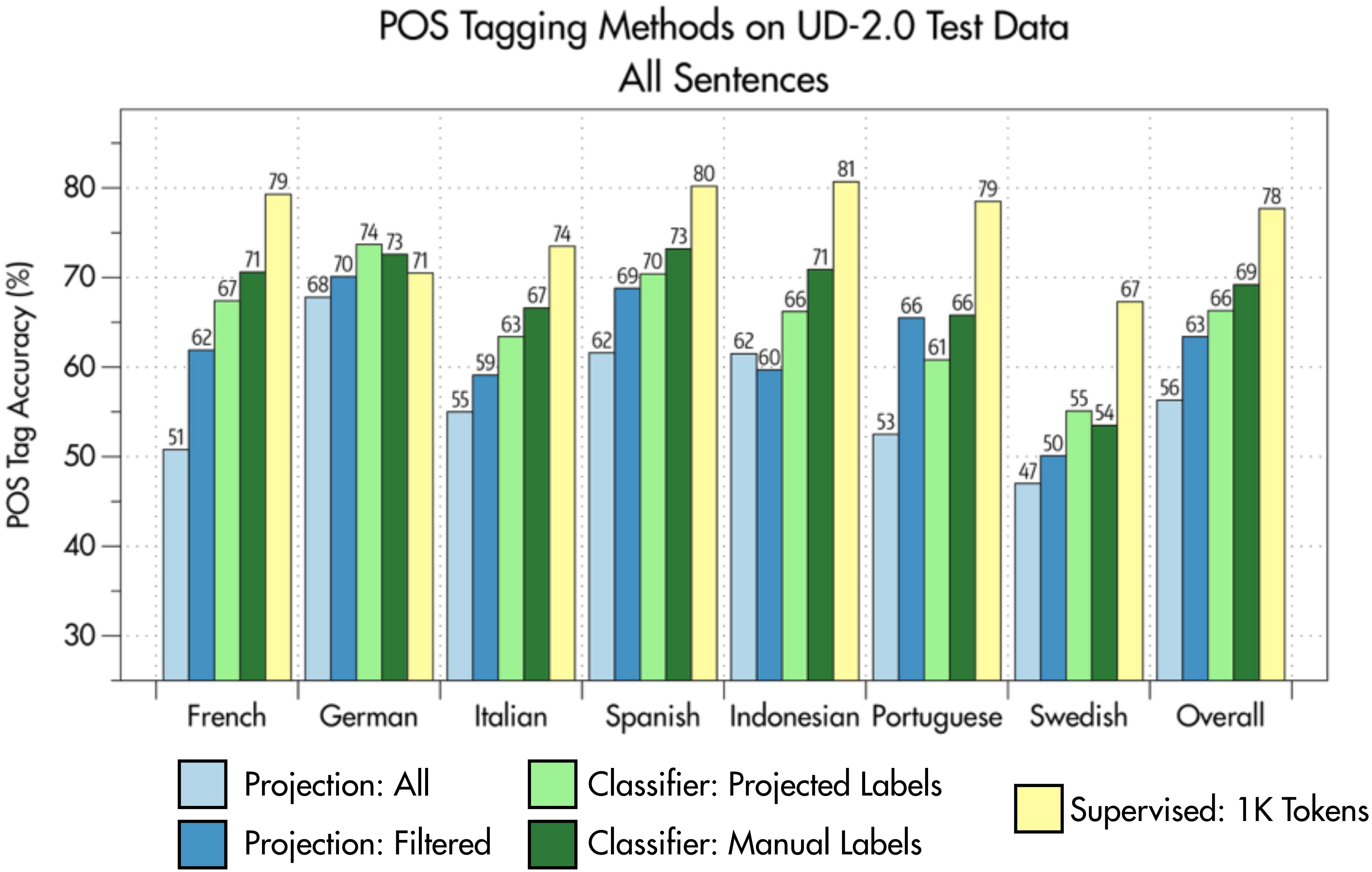
Monolingual POS Tagging

- Four settings:
  - Projection
    - **All** instances: **with** unaligned words
    - **Filtered** instances: **no** unaligned words
  - Classification:
    - **Projected** training tokens

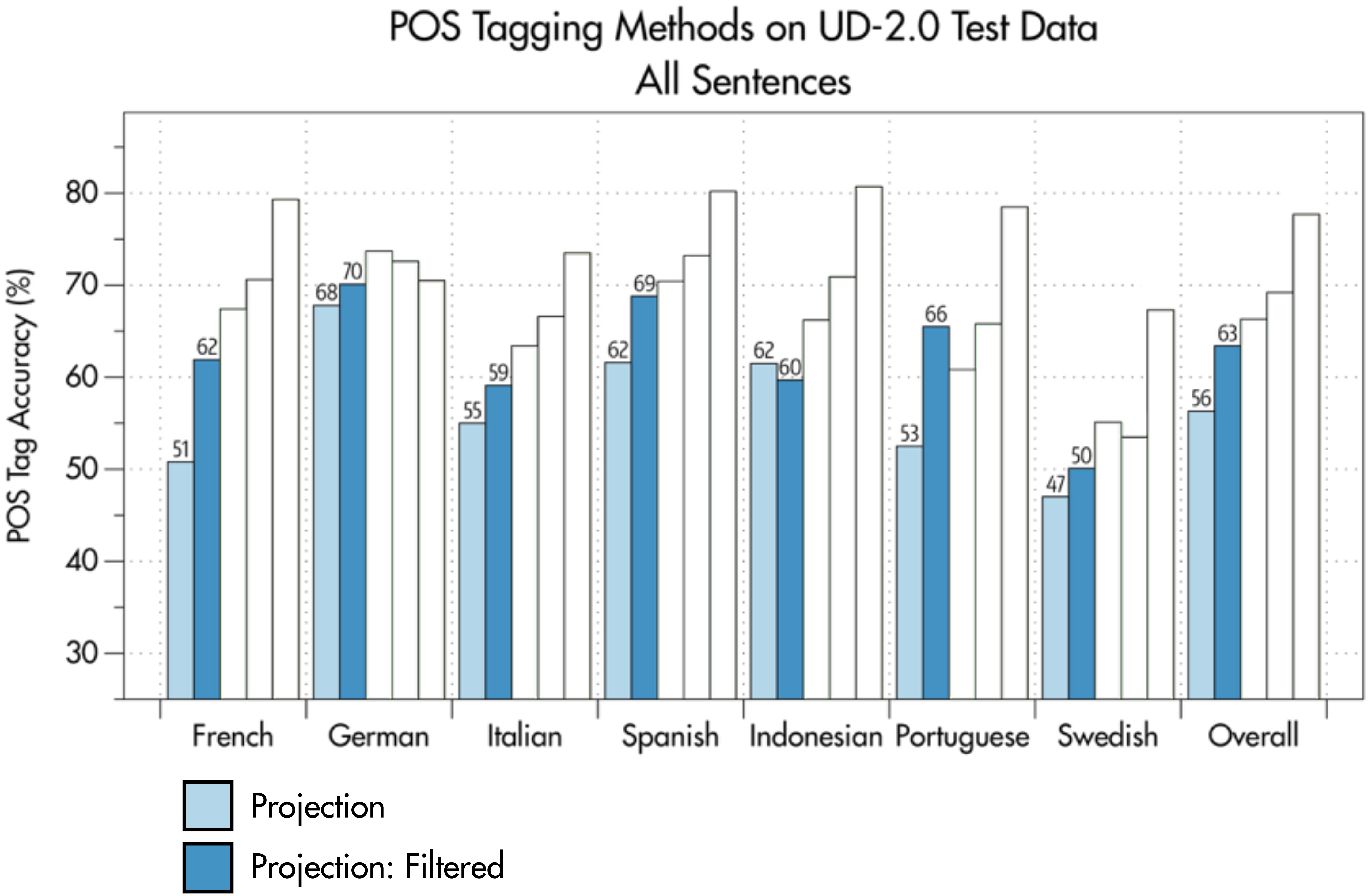
Monolingual POS Tagging

- Four settings:
  - Projection
    - **All** instances: **with** unaligned words
    - **Filtered** instances: **no** unaligned words
  - Classification:
    - **Projected** training tokens
    - **Manual** training tokens

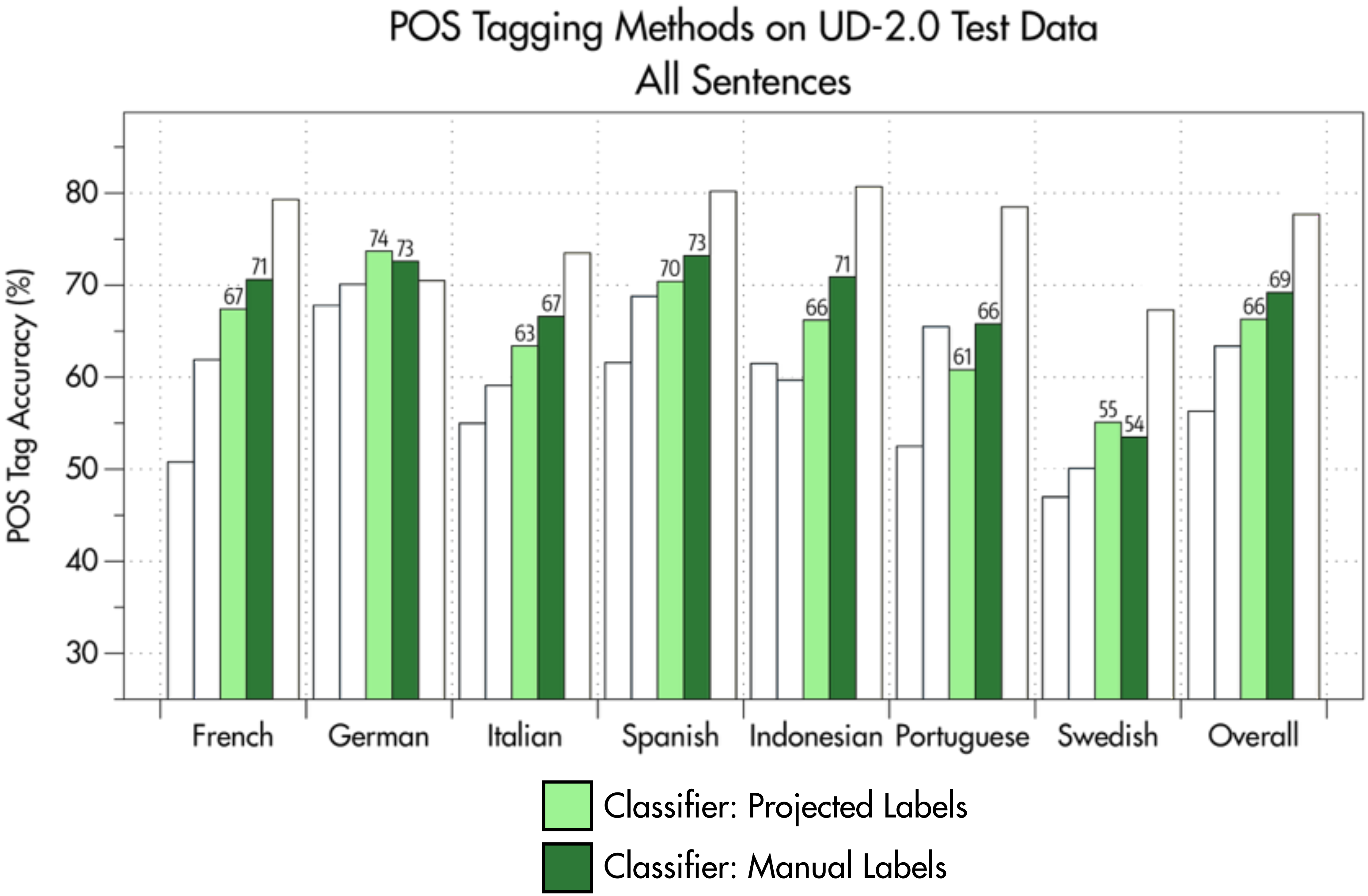
Monolingual POS Tagging



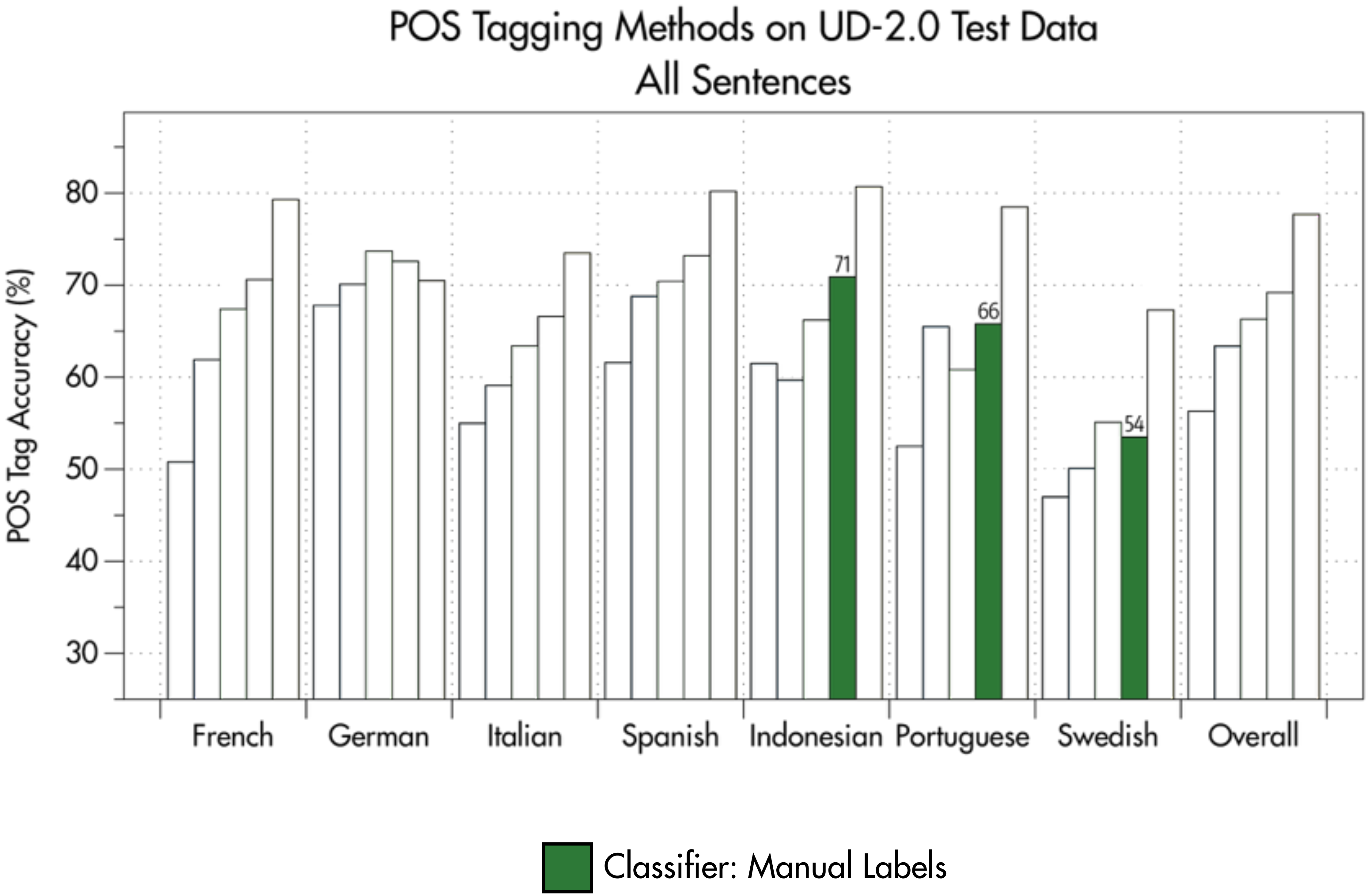
Monolingual POS Tagging



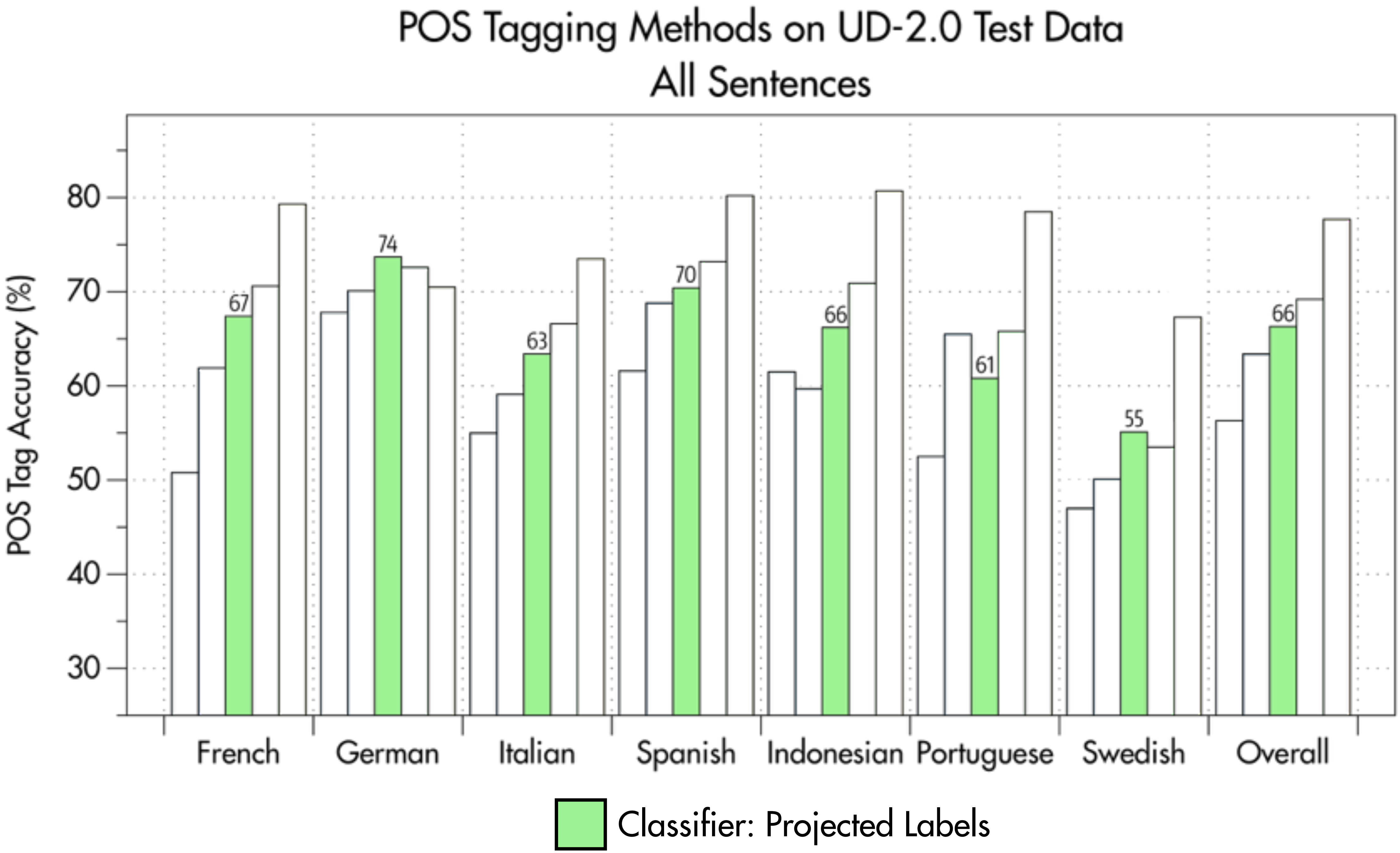
Monolingual POS Tagging



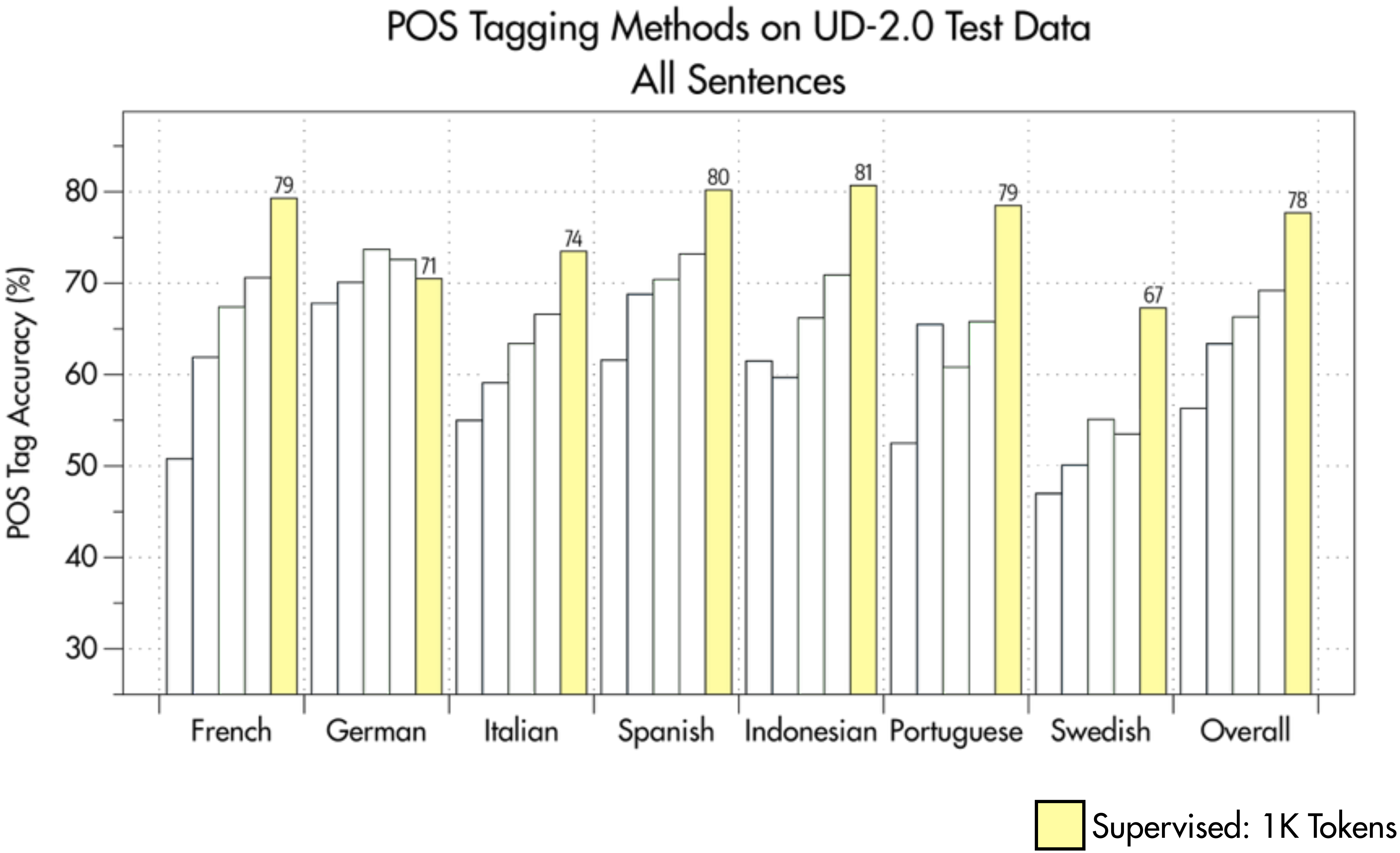
Monolingual POS Tagging



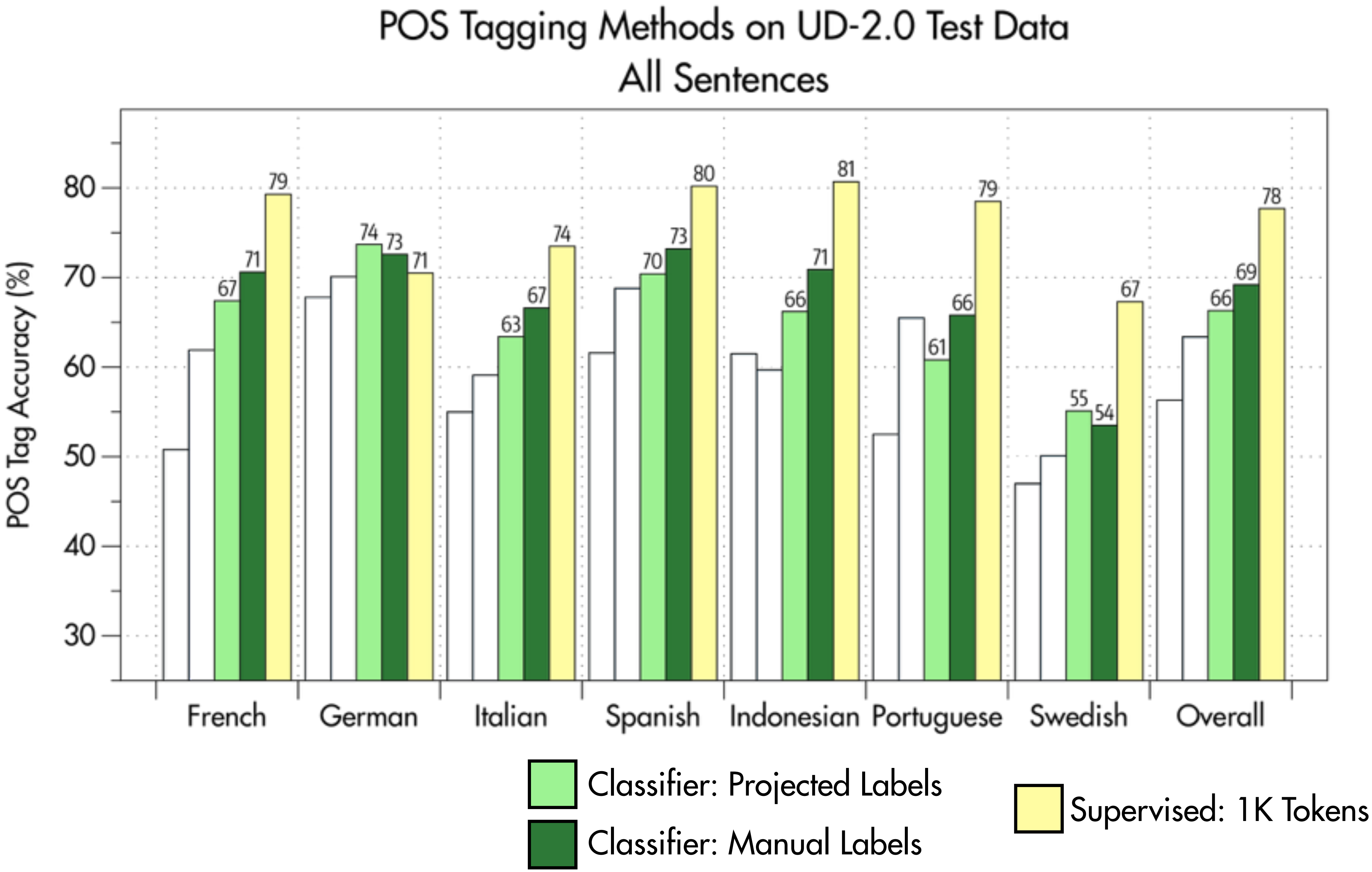
Monolingual POS Tagging



Monolingual POS Tagging



Monolingual POS Tagging



Monolingual POS Tagging

- Another variable in using the UD-2.0 corpus:
  - Corpus represents a large shift in domain

Monolingual POS Tagging

- Another variable in using the UD-2.0 corpus:
  - Corpus represents a large shift in domain
  - UD-2.0:
    - 20.8 words/sentence
    - Newswire

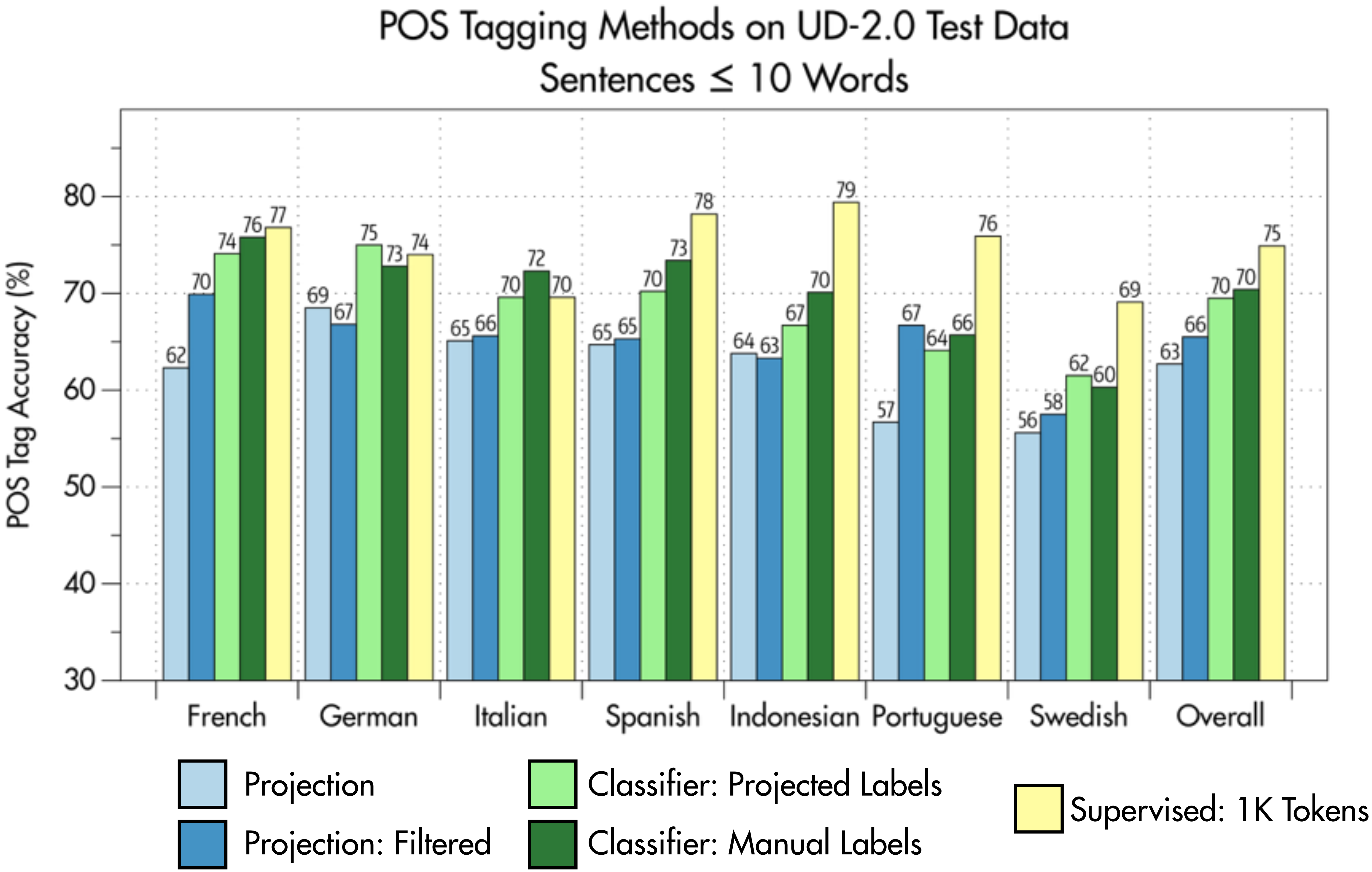
Monolingual POS Tagging

- Another variable in using the UD-2.0 corpus:
  - Corpus represents a large shift in domain
  - UD-2.0:
    - 20.8 words/sentence
    - Newswire
  - IGT sentences
    - 6.1 words/instance
    - Illustrative examples

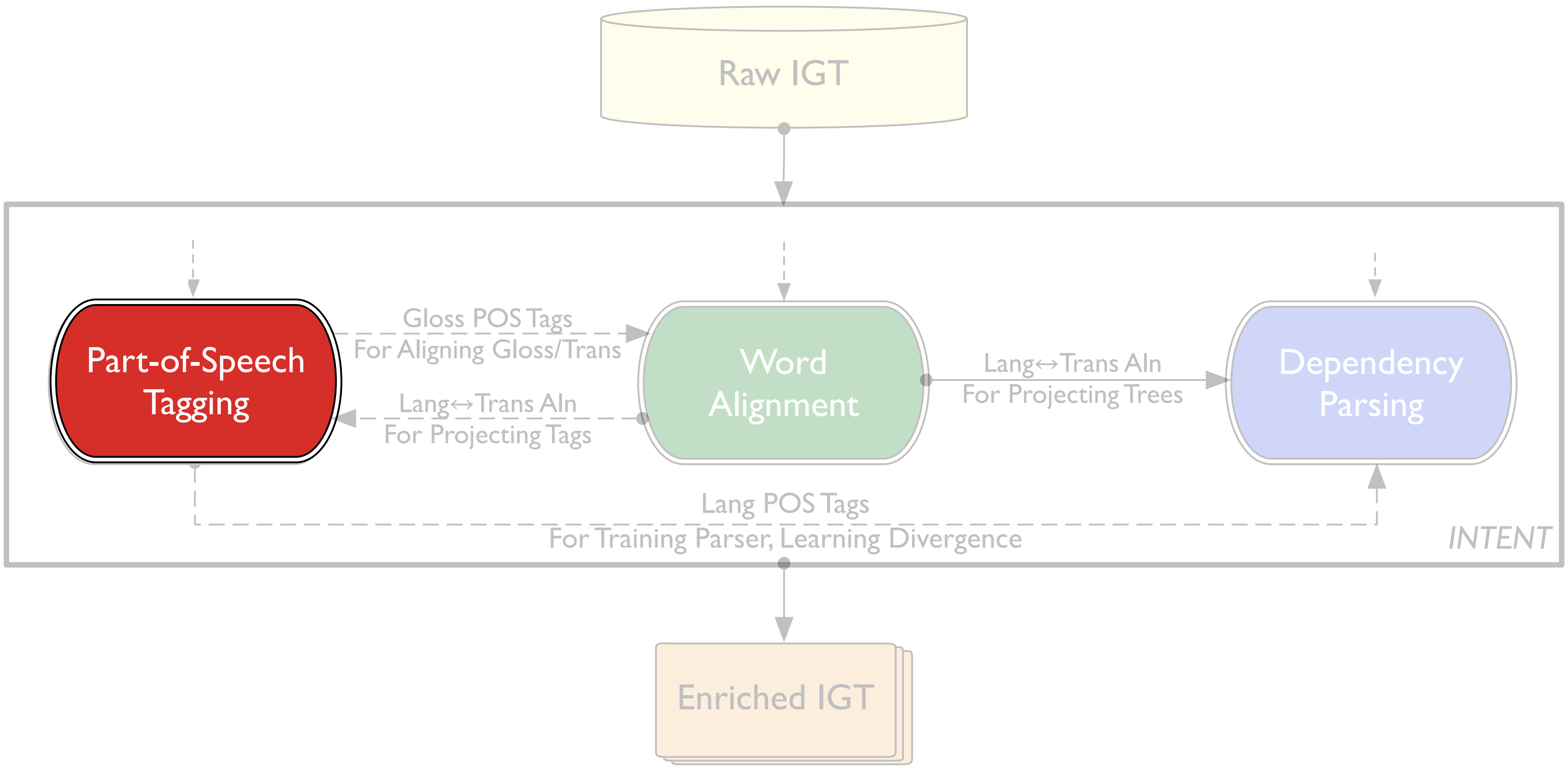
Monolingual POS Tagging

- Another variable in using the UD-2.0 corpus:
  - Corpus represents a large shift in domain
  - UD-2.0:
    - 20.8 words/sentence
    - Newswire
  - IGT sentences
    - 6.1 words/instance
    - Illustrative examples
- Try evaluating also on short UD-2.0 sentences

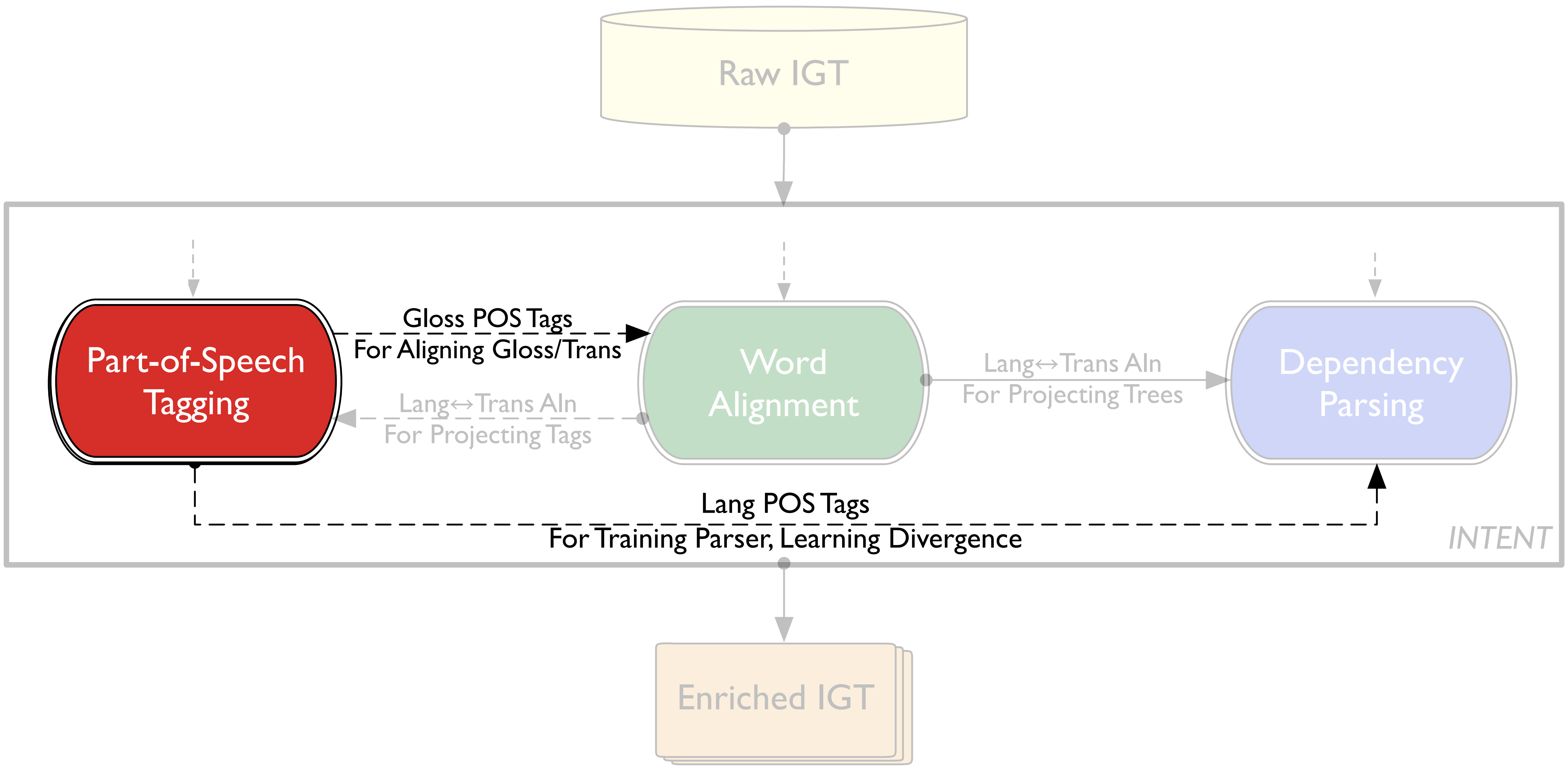
Monolingual POS Tagging



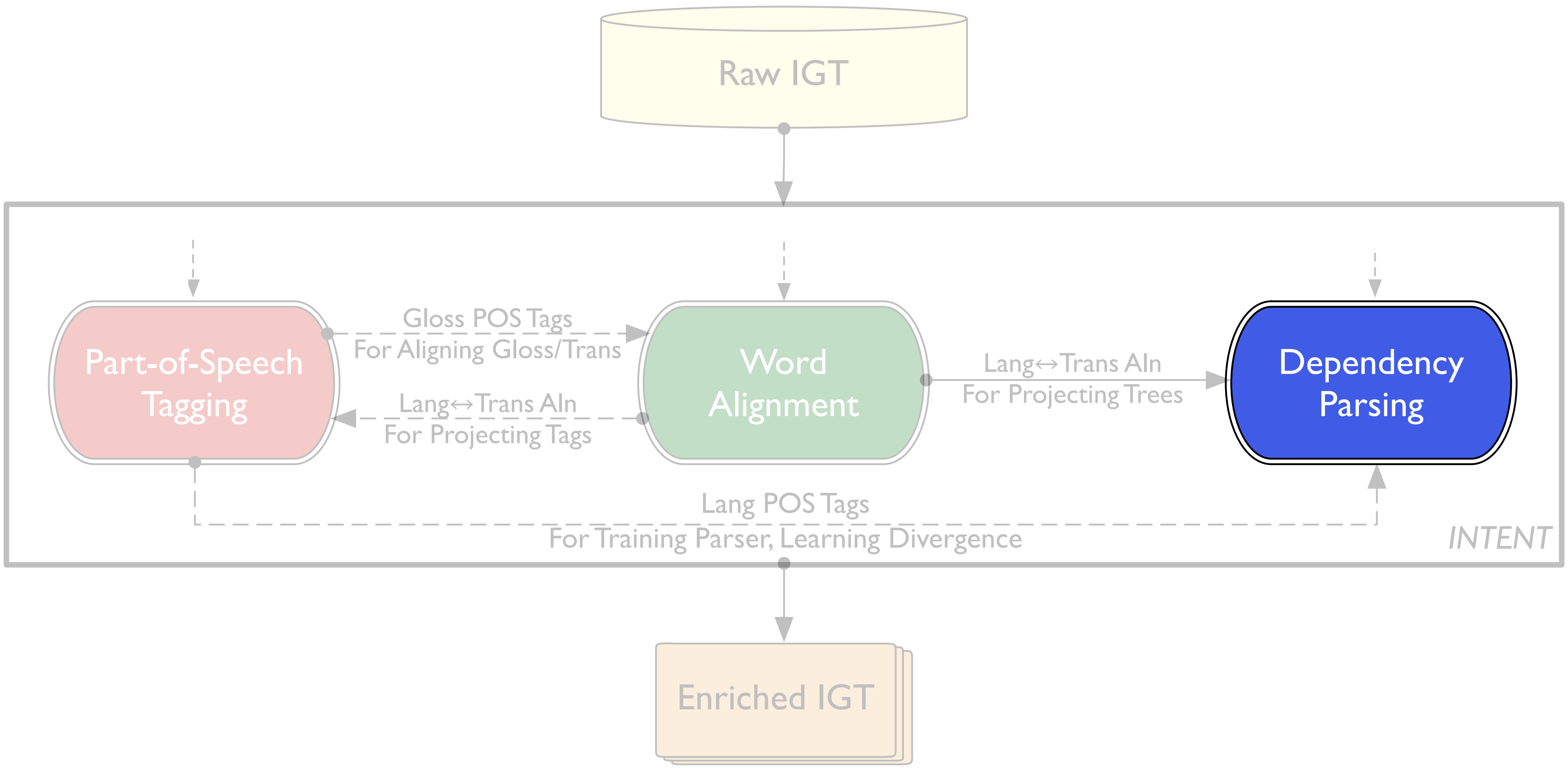
The INTENT System



The INTENT System



The INTENT System



DS Projection

The teacher gave a book to the boy yesterday

Rhoddod yr athro lyfr i'r bachgen ddoe

Welsh *[cym] – (Bailyn, 2004)*

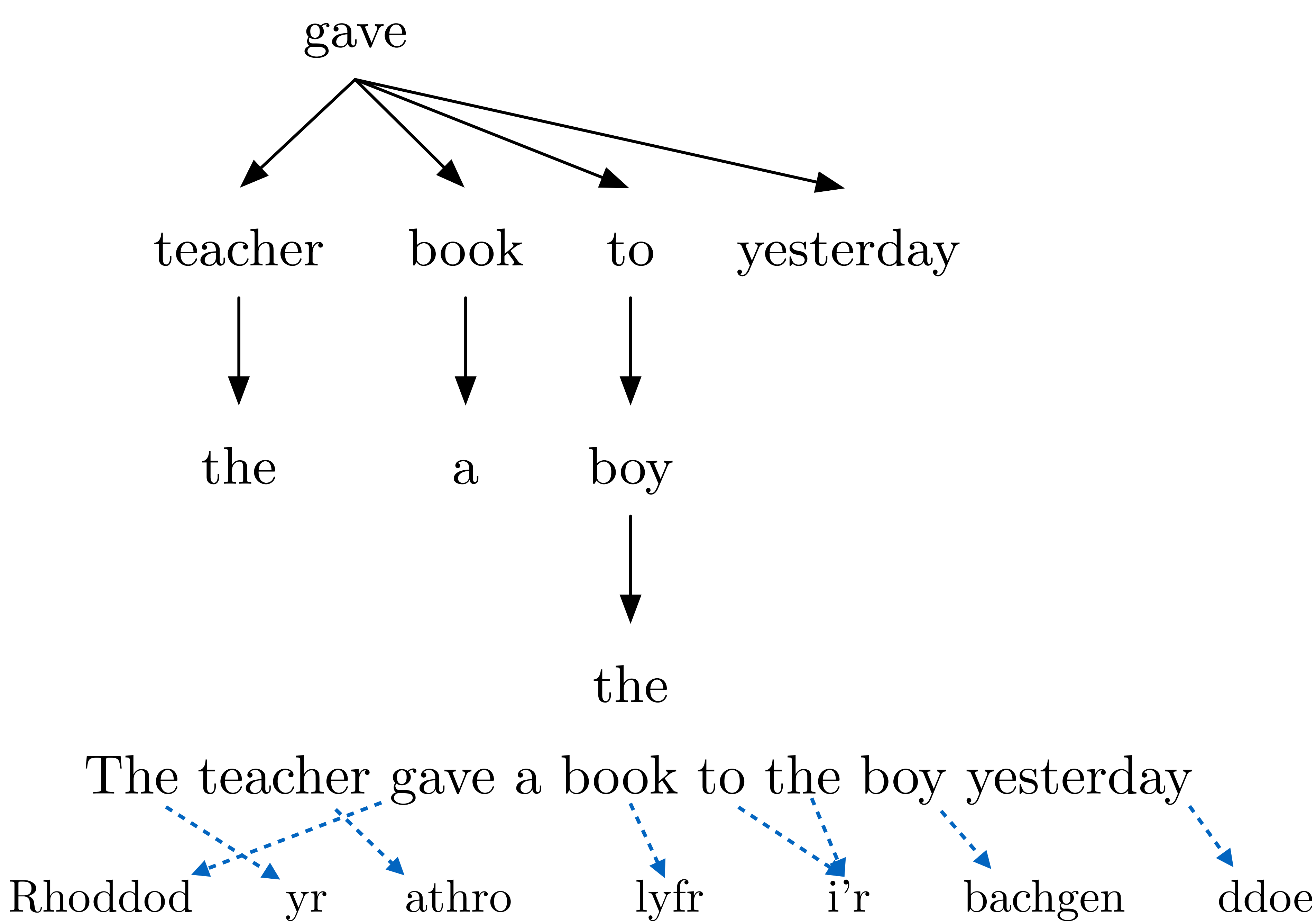
DS Projection

The teacher gave a book to the boy yesterday

Rhoddod yr athro lyfr i'r bachgen ddoe

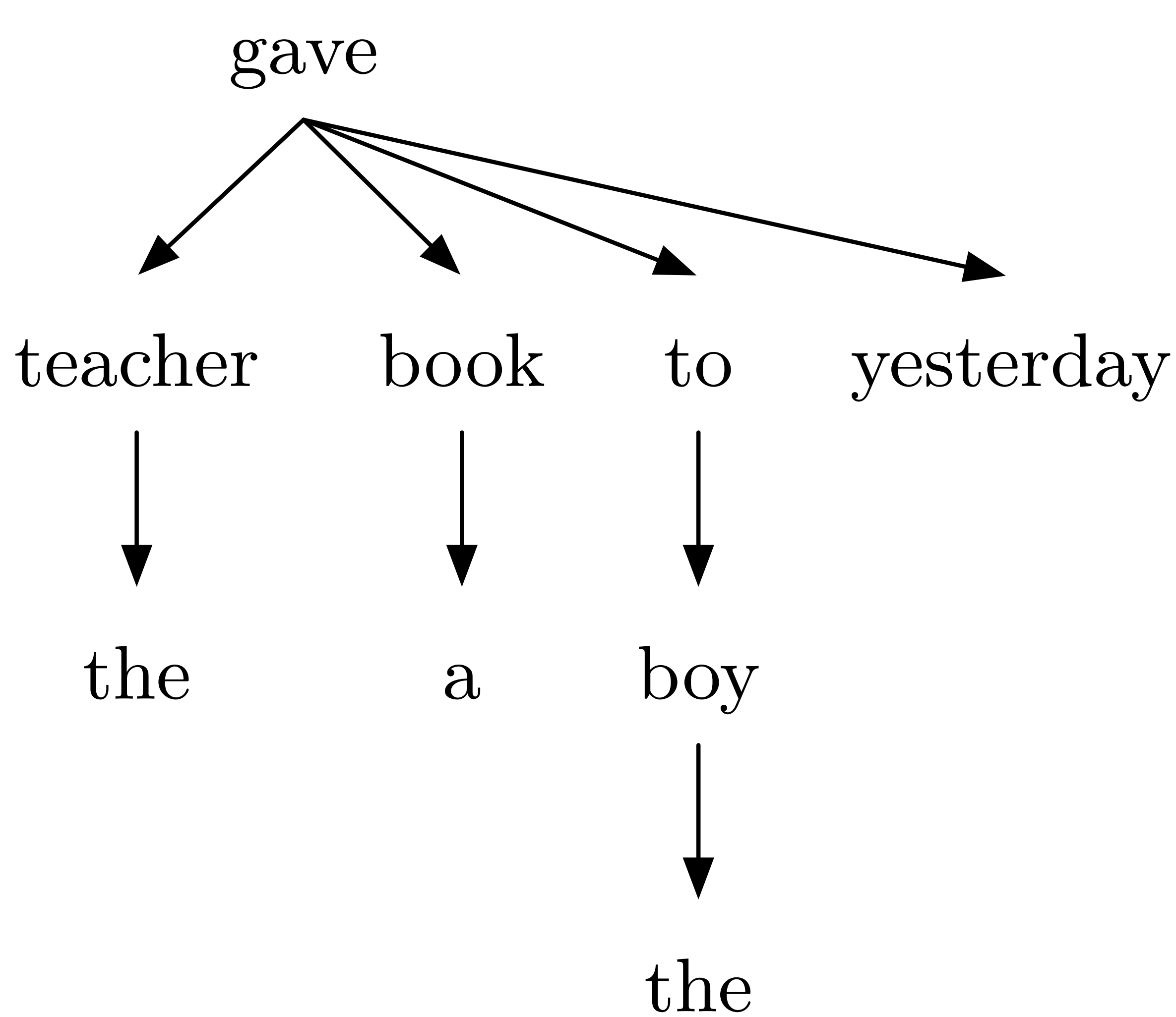
Welsh *[cym]* – *(Bailyn, 2004)*

DS Projection



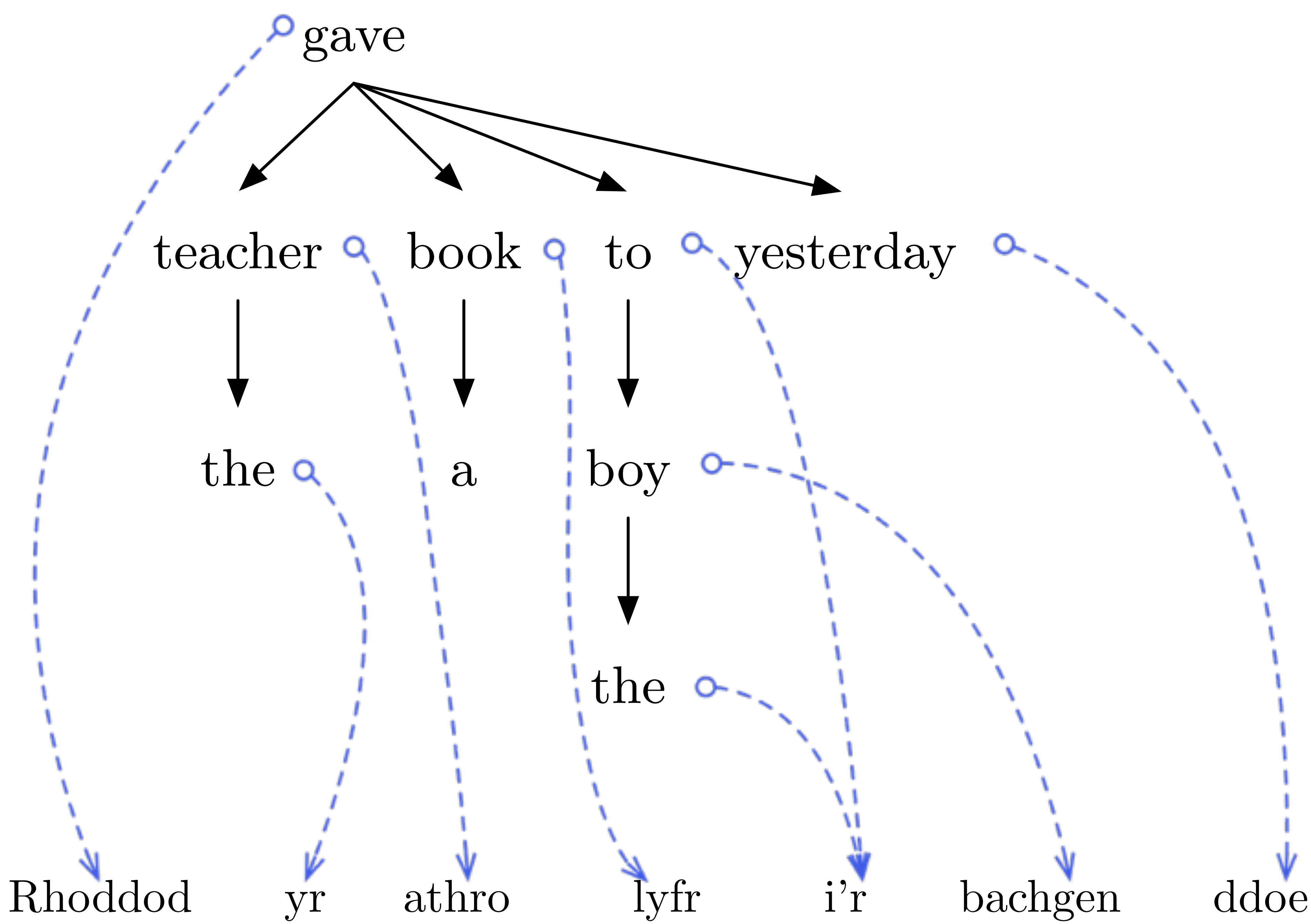
Welsh *[cym] – (Bailyn, 2004)*

DS Projection

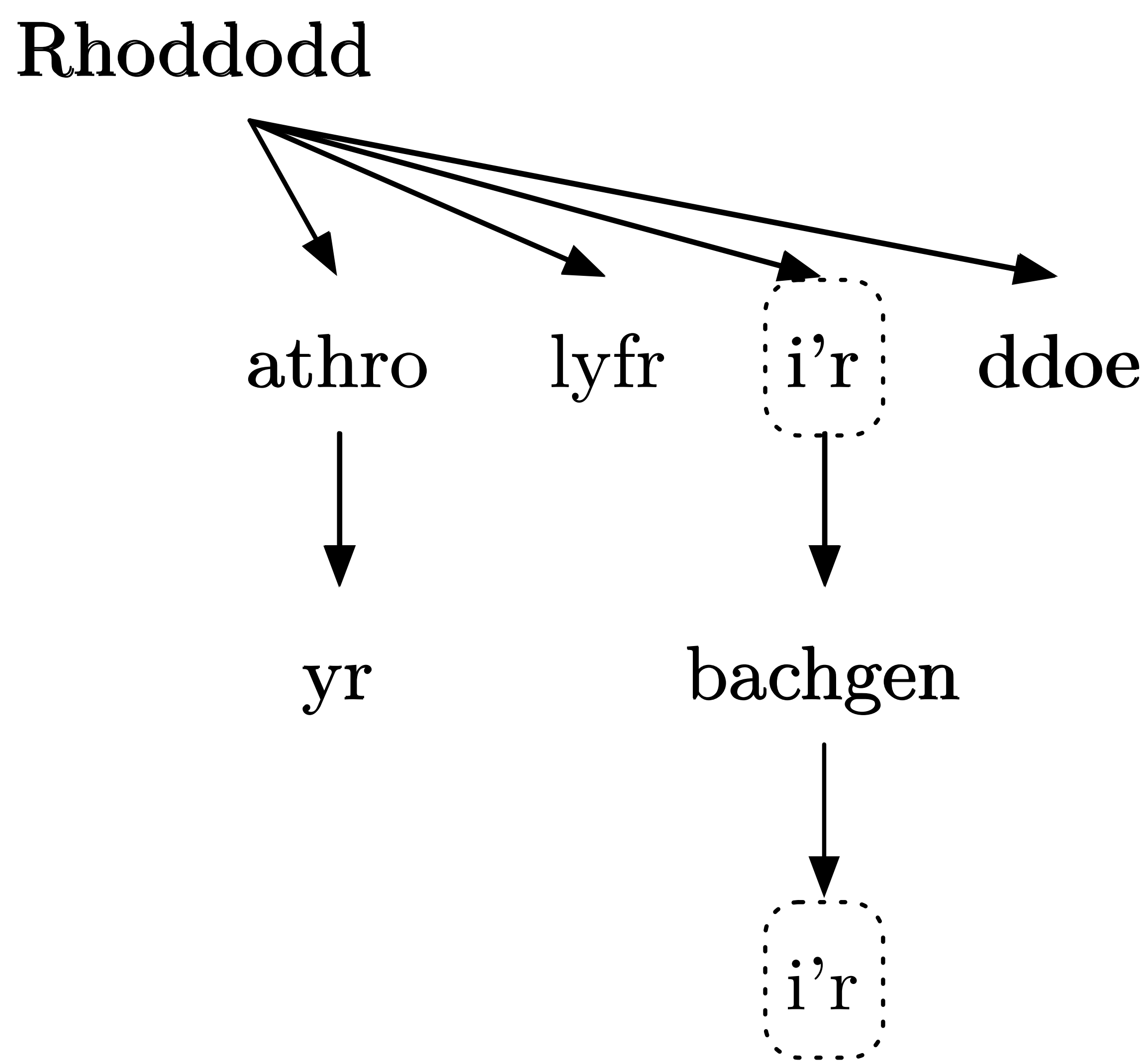


Rhoddod yr athro lyfr i'r bachgen ddoe

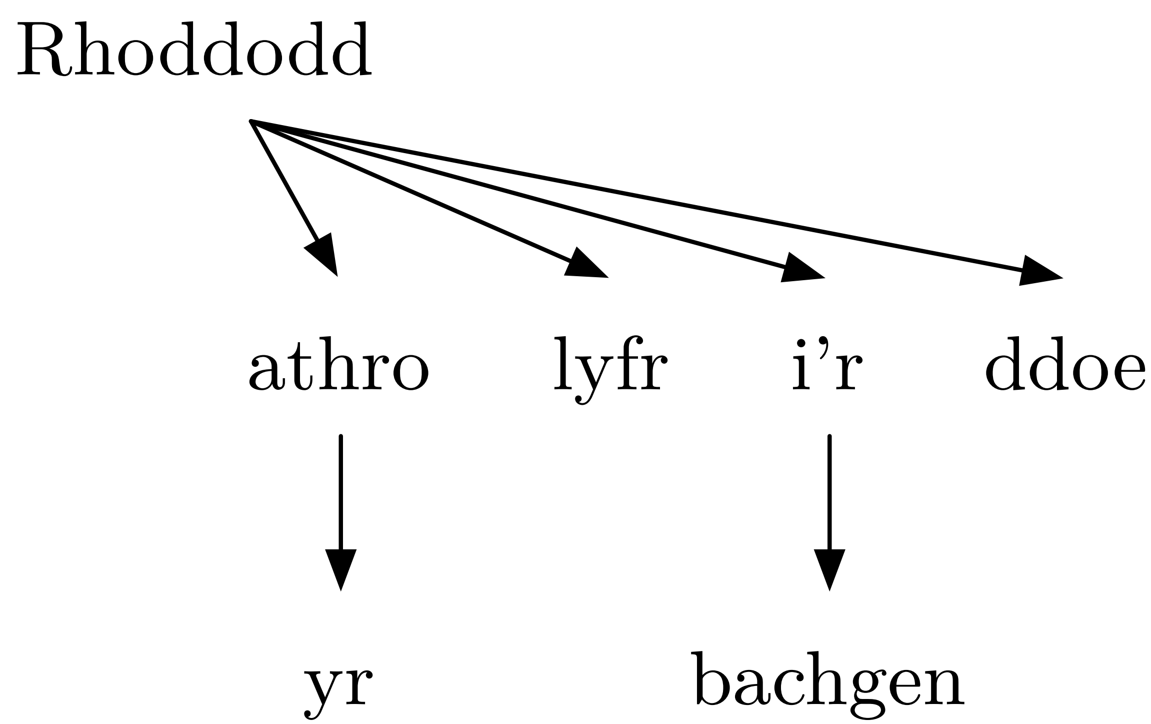
DS Projection



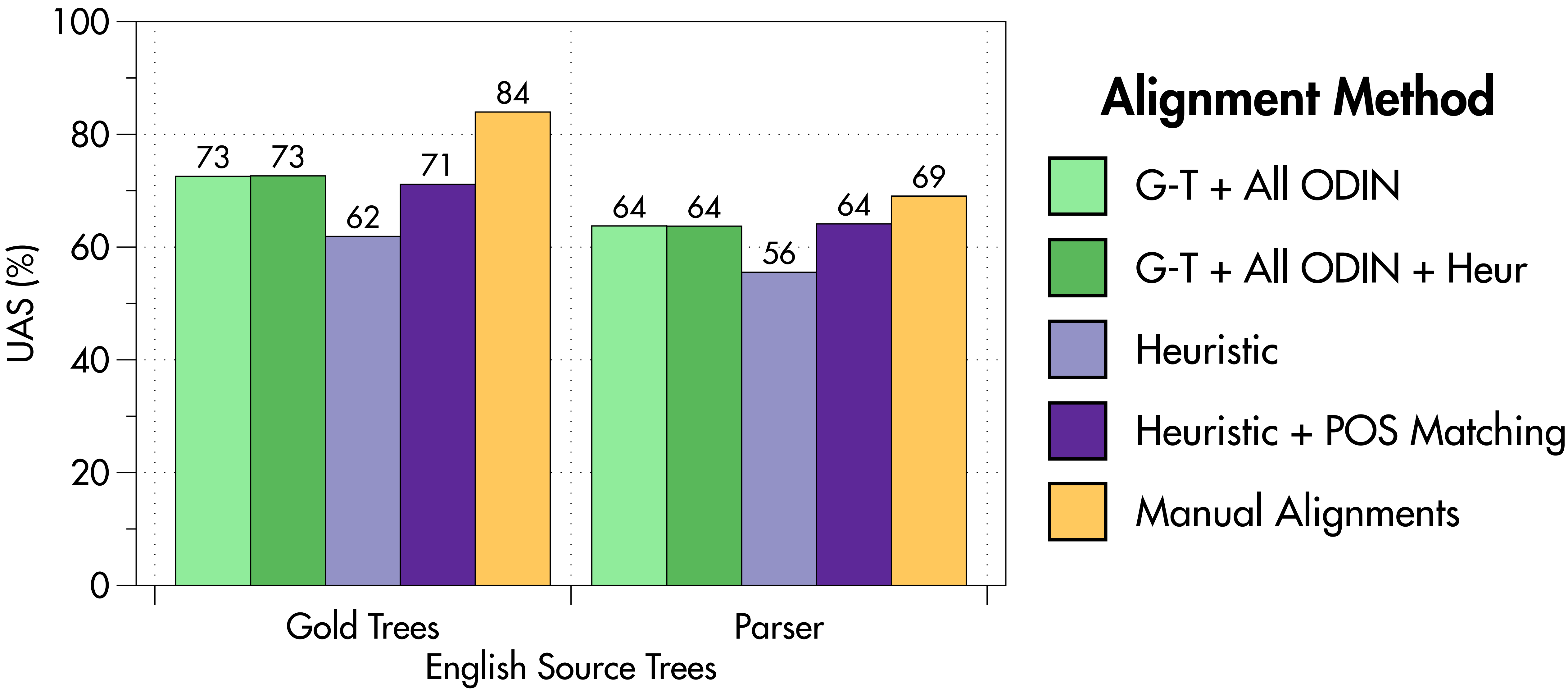
DS Projection



DS Projection



Dependency Parsing: Projection



Language Divergence

Language Divergence

- Direct Correspondence Assumption (DCA) (Hwa et. al, 2005)

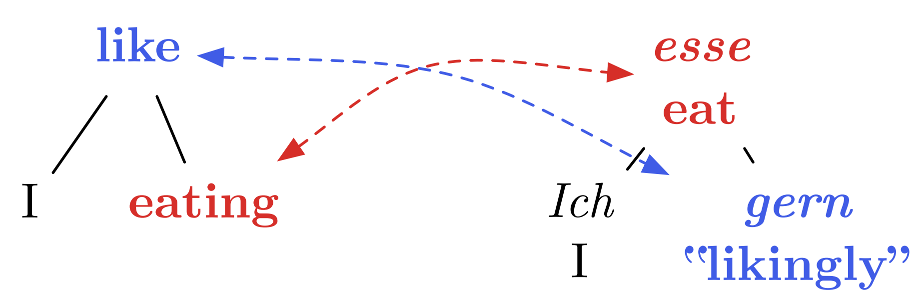
Language Divergence

- Direct Correspondence Assumption (DCA) (Hwa et. al, 2005)
- Language Divergence (Dorr, 1994)

Divergence Types

Head-Switching Divergence

English	**German**
I like eating	*Ich esse gern*
	("I eat likingly')



Promotional Divergence

Divergence Types

Structural Divergence

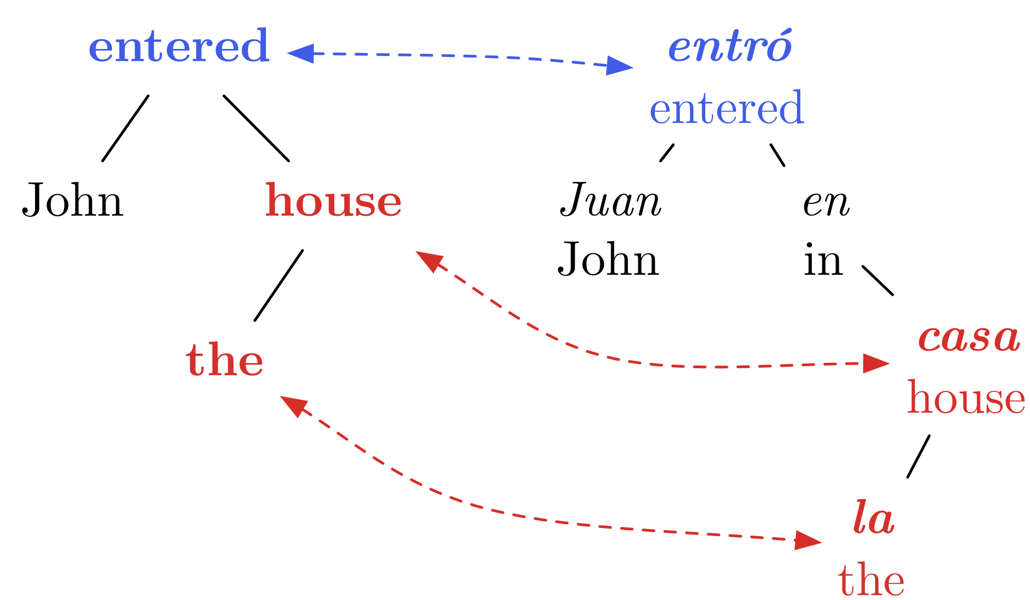
English

John entered the house

Spanish

Juan entró en la casa

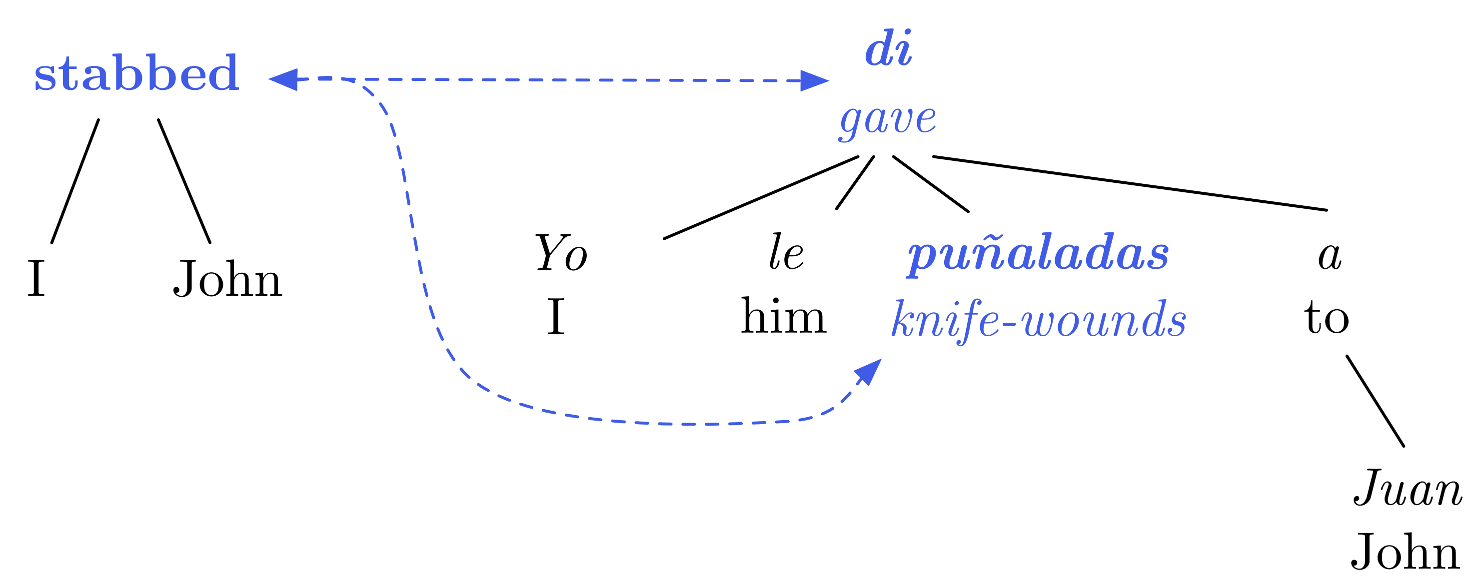
("John entered in the house")



Divergence Types

Conflational Divergence

English	**Spanish**
I stabbed John	*Yo le di puñaladas a Juan*
	("I gave knife-wounds to John")

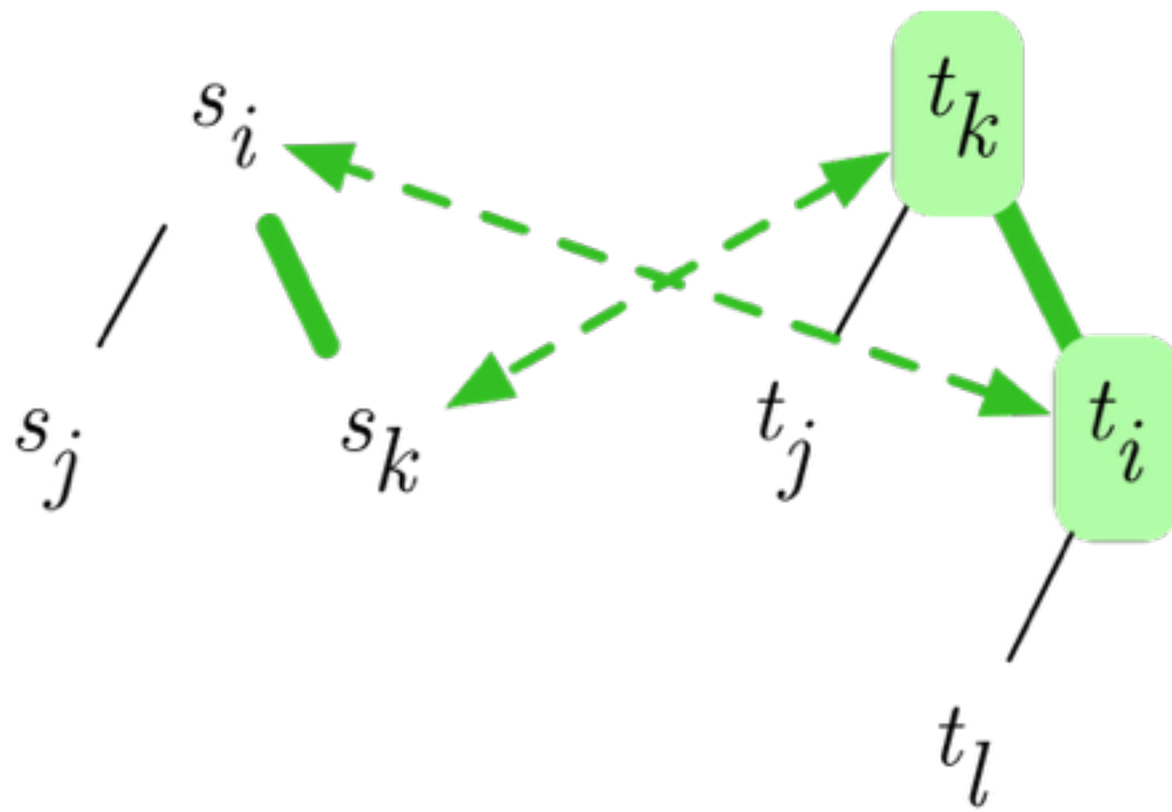


Addressing DS Divergence

- Results for DS projection on IGT show divergence
- Learn when projection is unreliable?

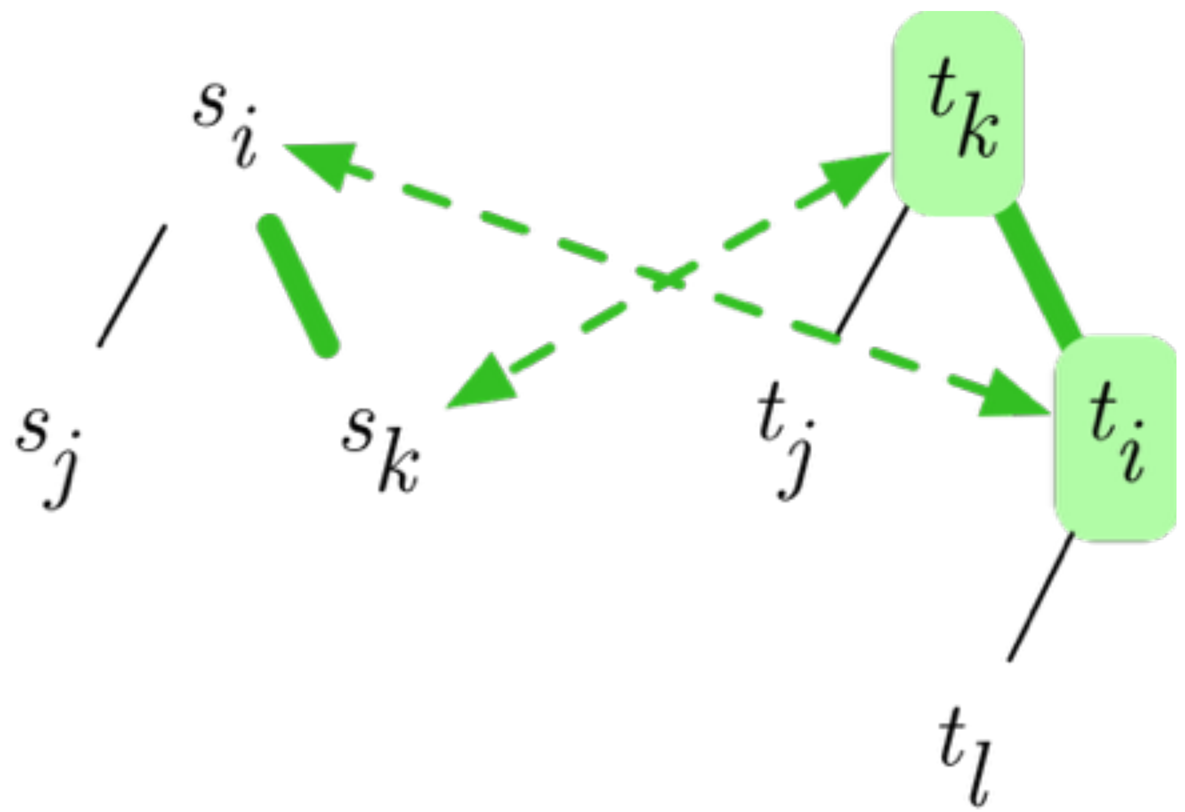
Alignment Types

Swap



Alignment Types

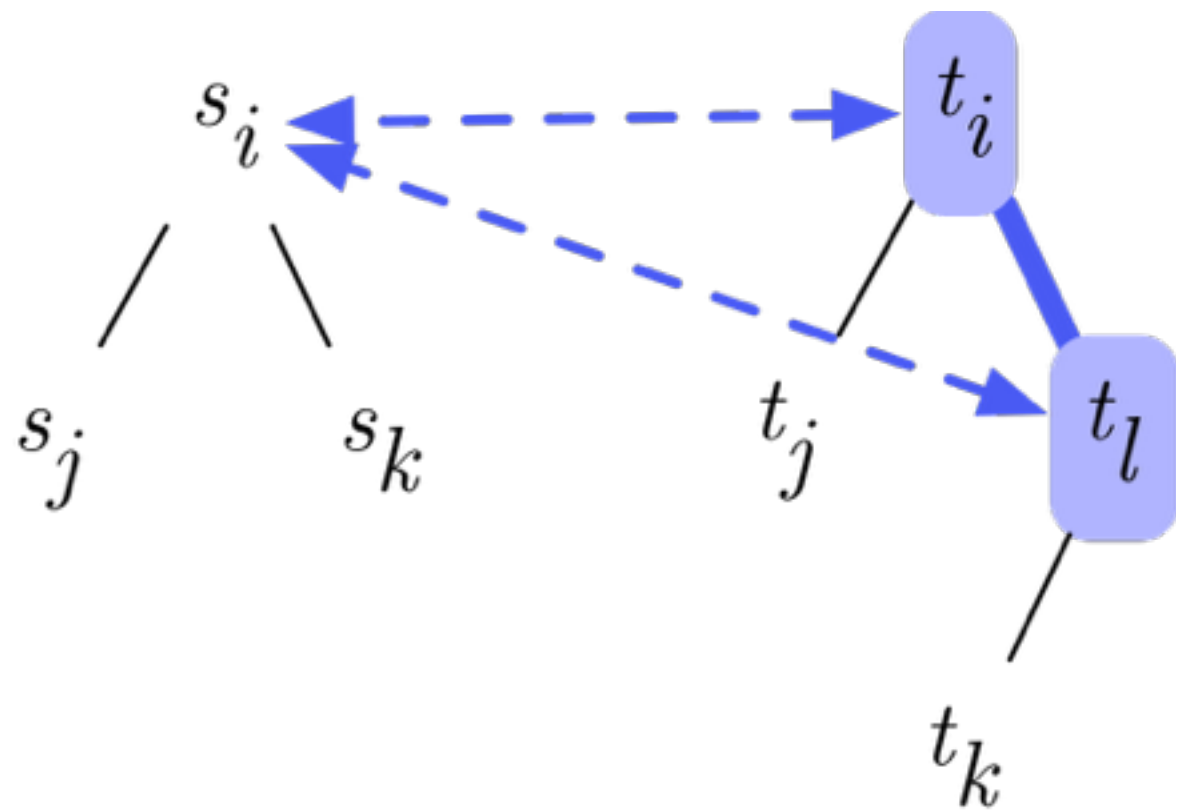
Swap



(Addresses **Head-Switching**)

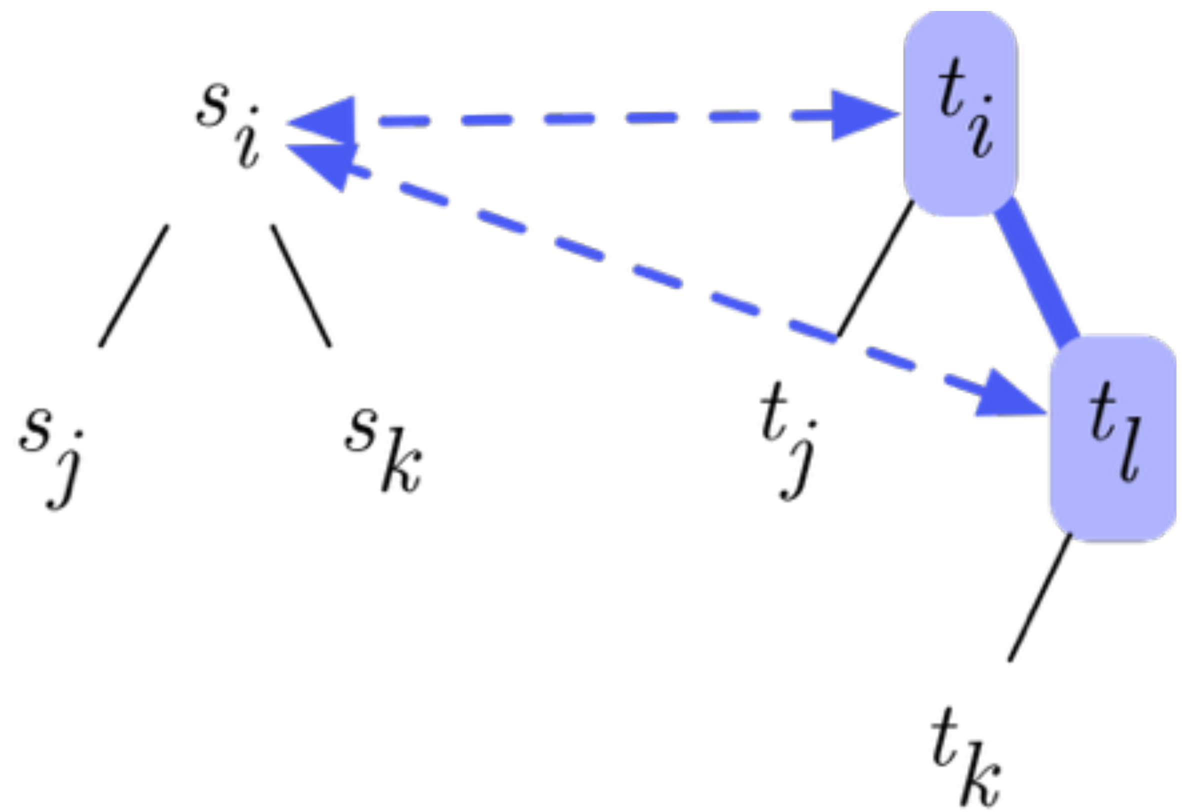
Alignment Types

Merge



Alignment Types

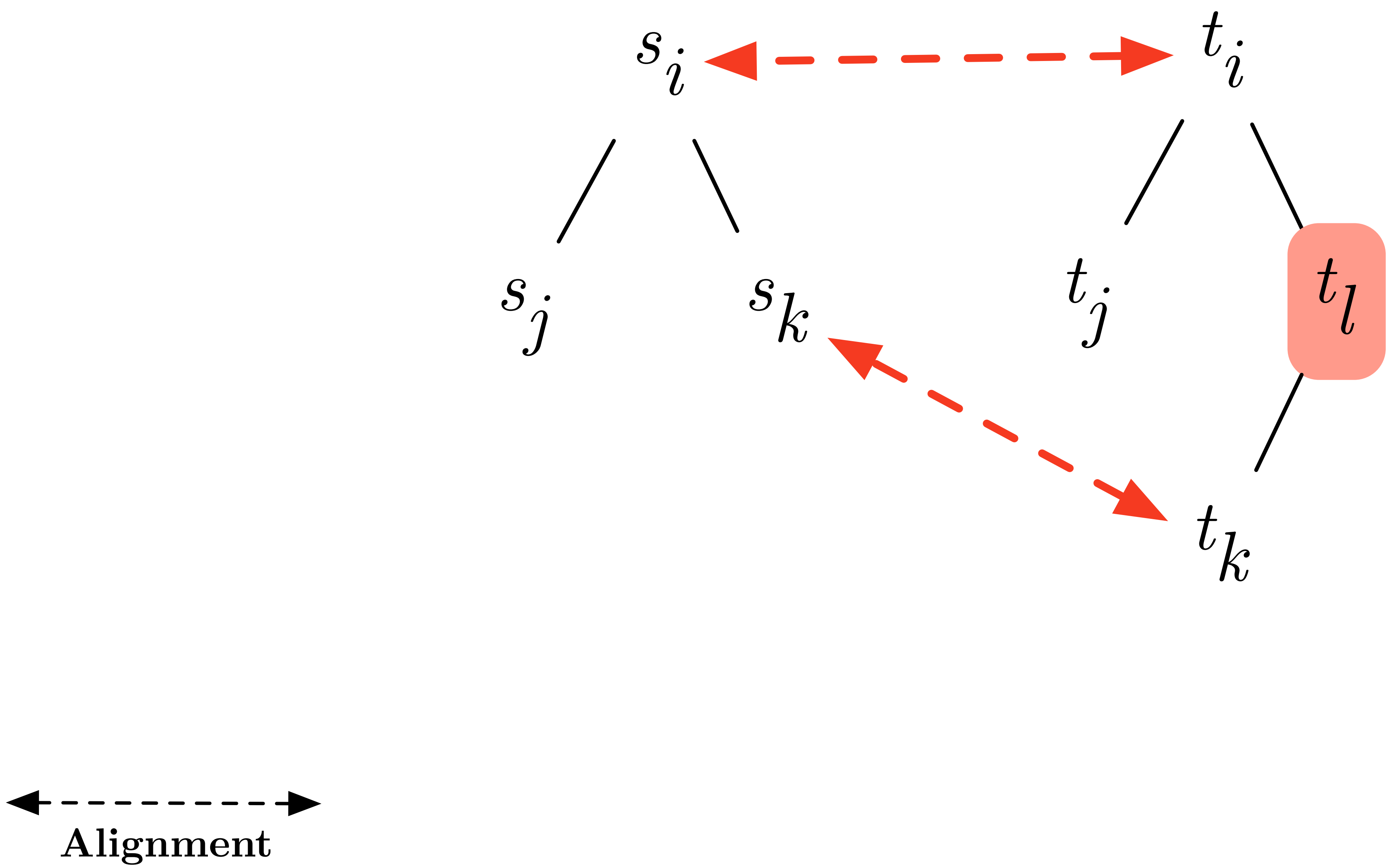
Merge



(Addresses **Conflational** Divergence)

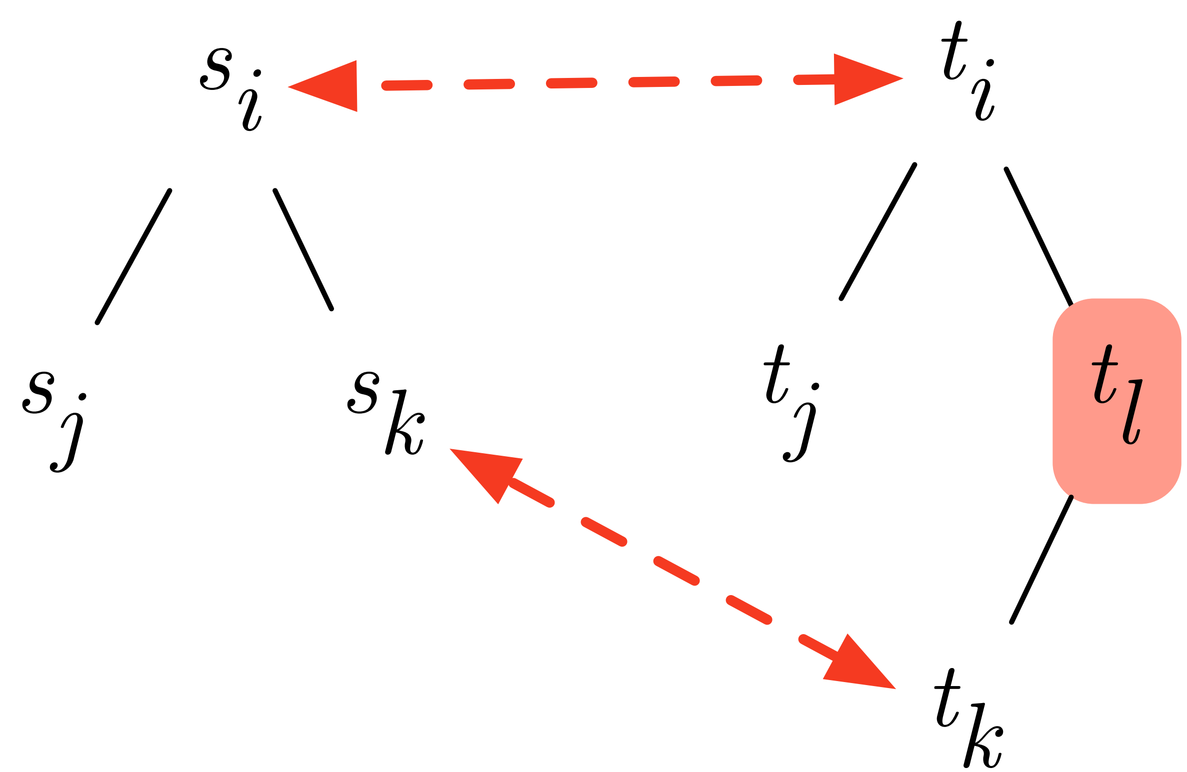
Alignment Types

Spontaneous

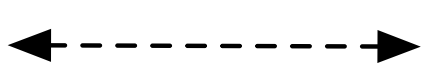


Alignment Types

Spontaneous

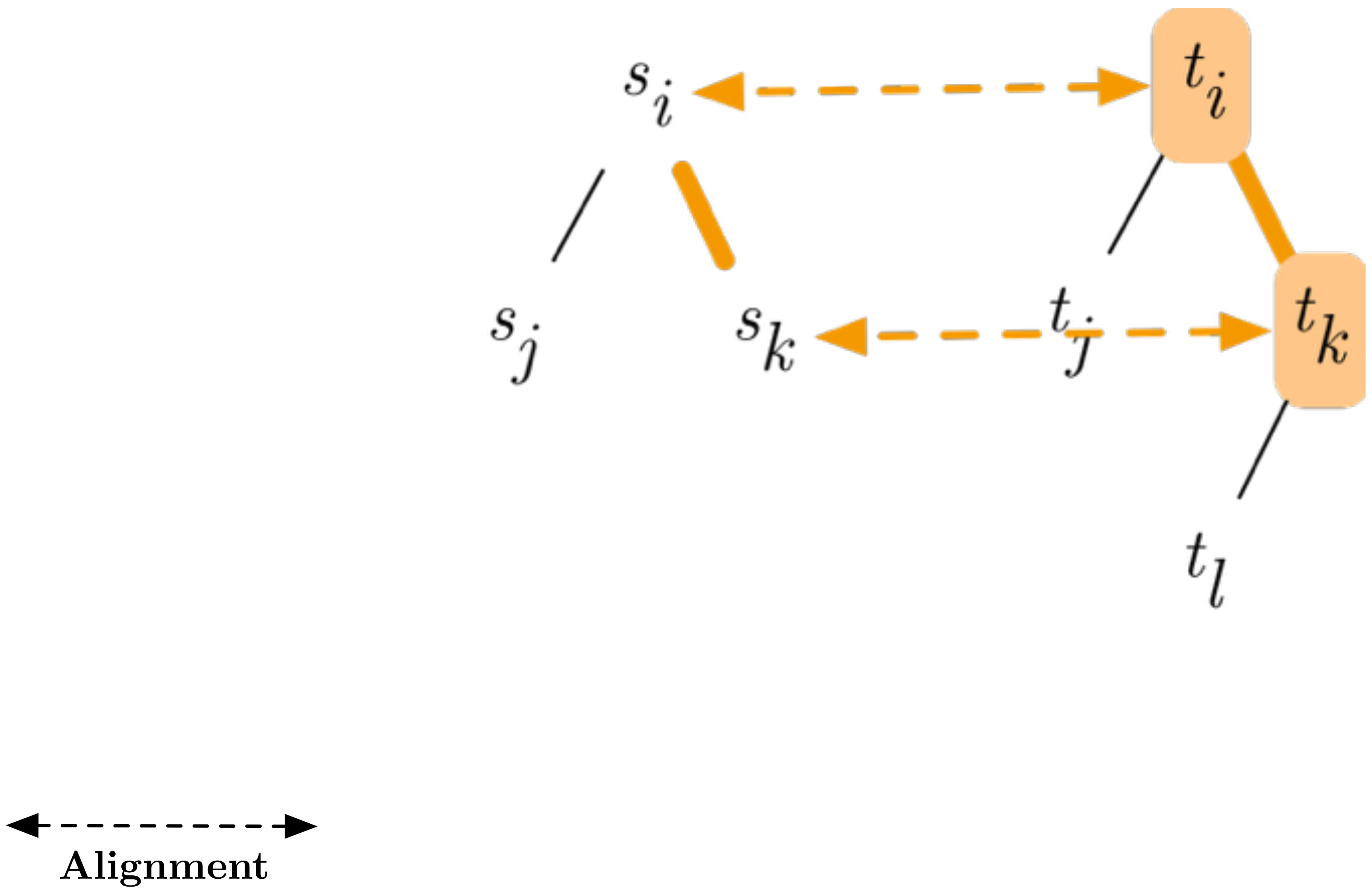


(Addresses **Structural** Divergence)



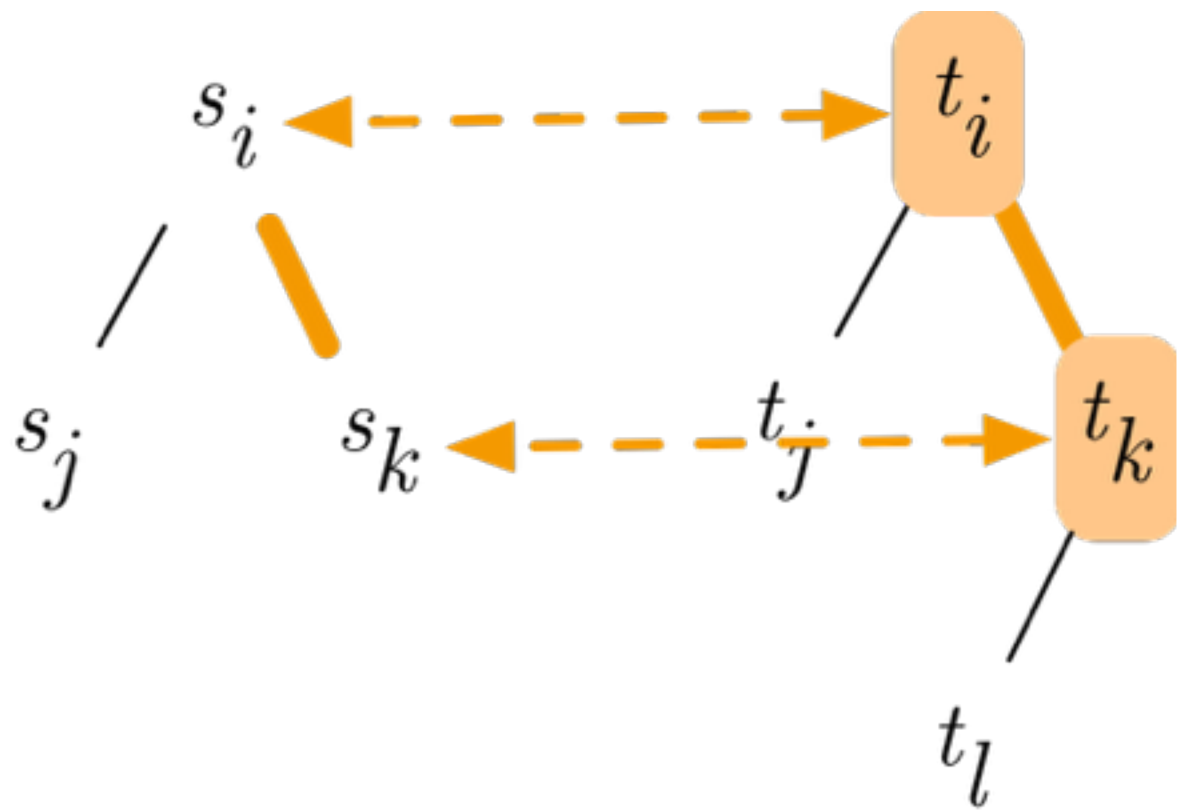
Alignment Types

Match



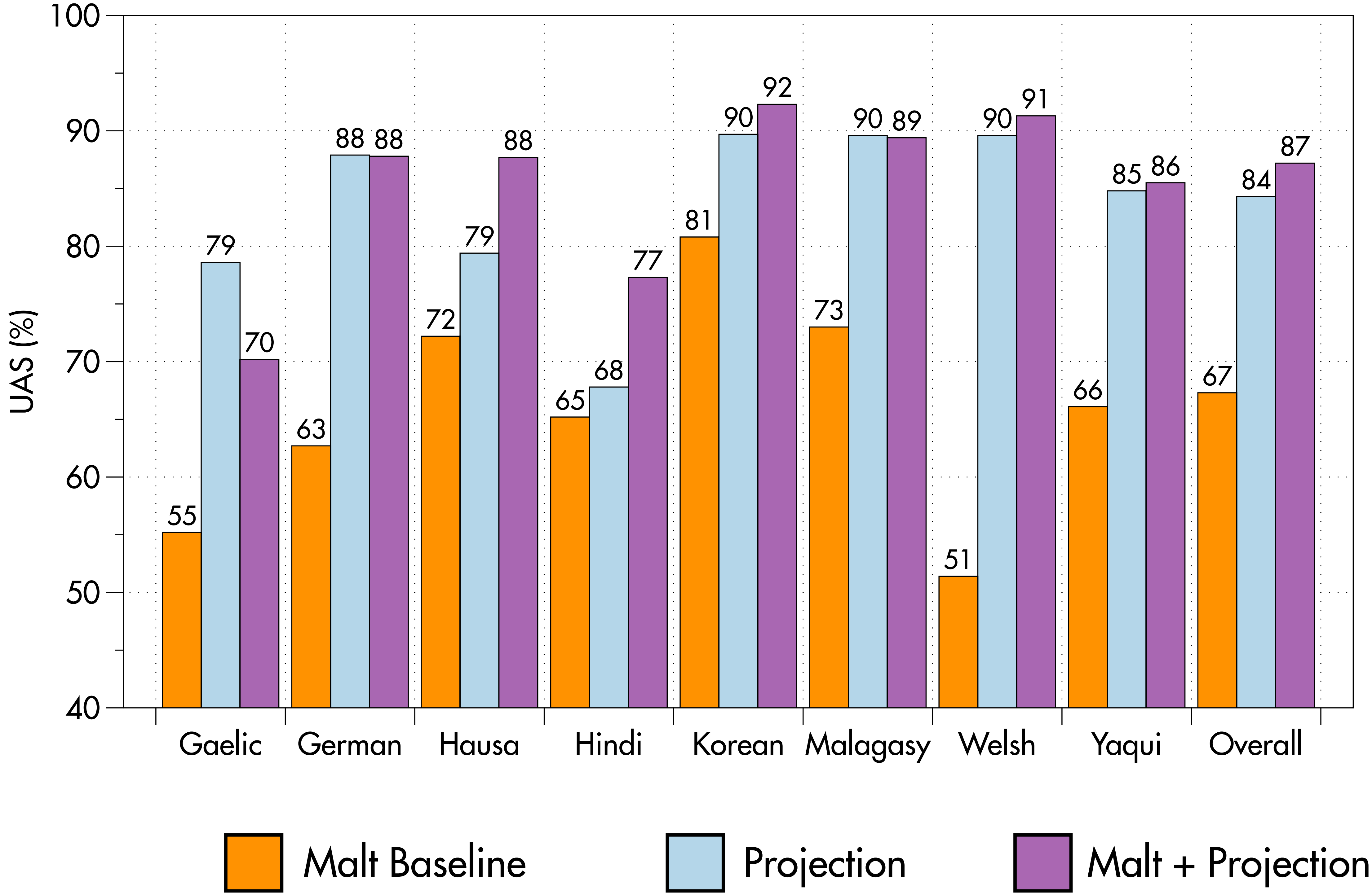
Alignment Types

Match



(**No** Divergence)

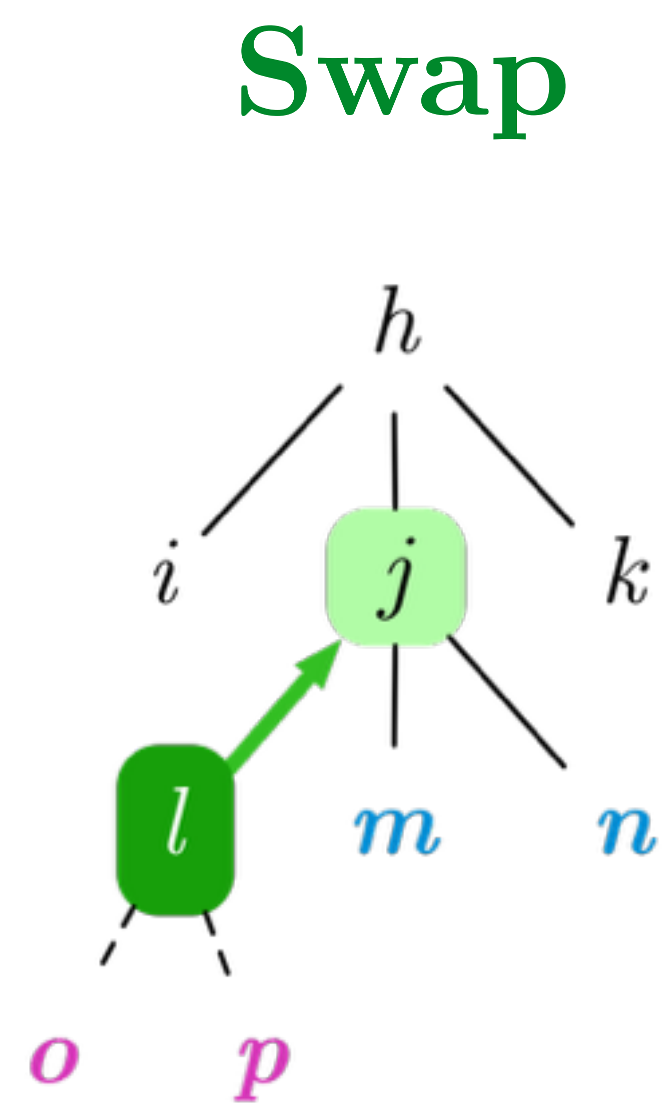
Projection-Enhanced Parsing



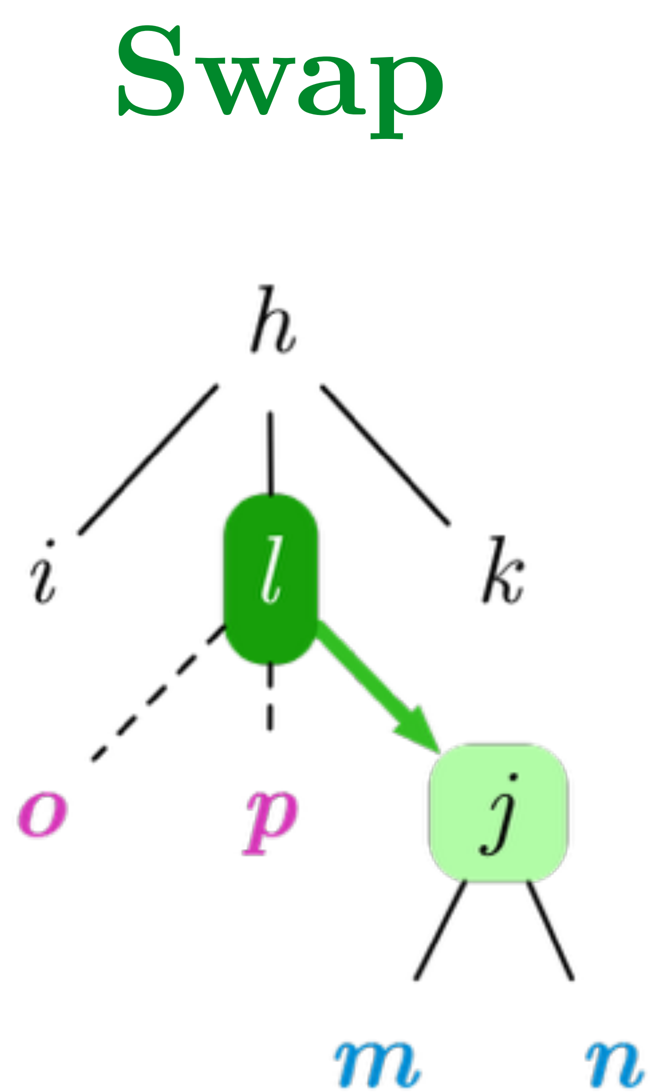
Measuring and "Correcting" Divergence

- Based on the idea of DUSTer (Dorr, 2002)
- Automatically rewrite dependency structures to pseudo-English that is more similar to target language

Tree Operations

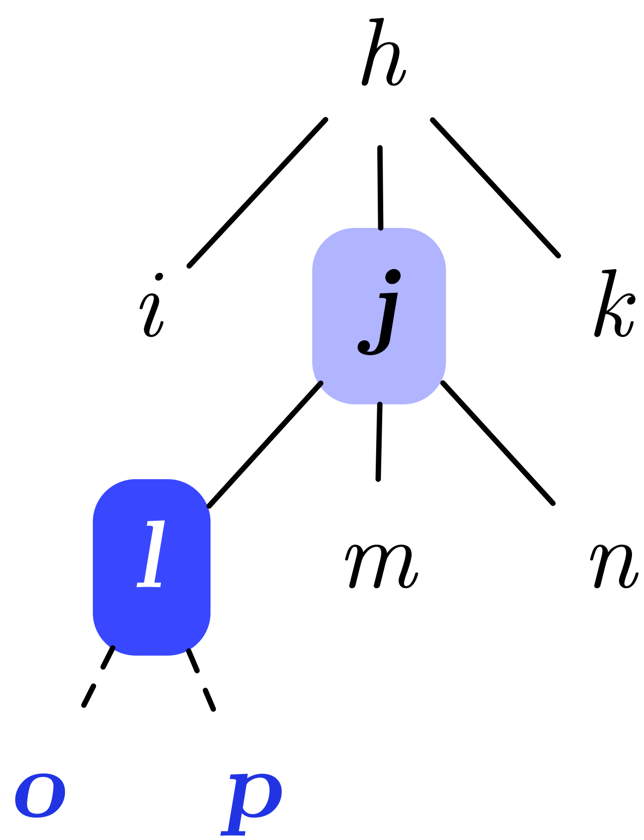


Tree Operations



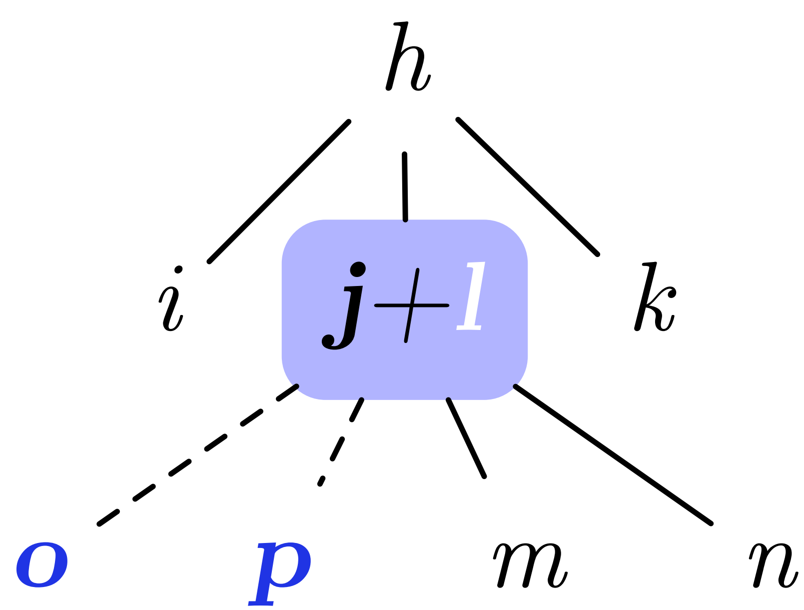
Tree Operations

Merge



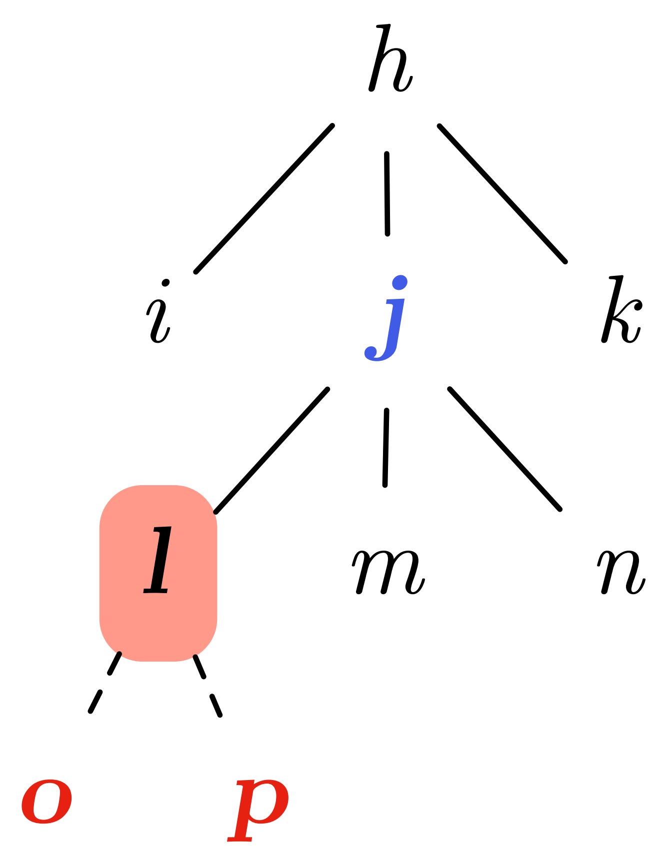
Tree Operations

Merge



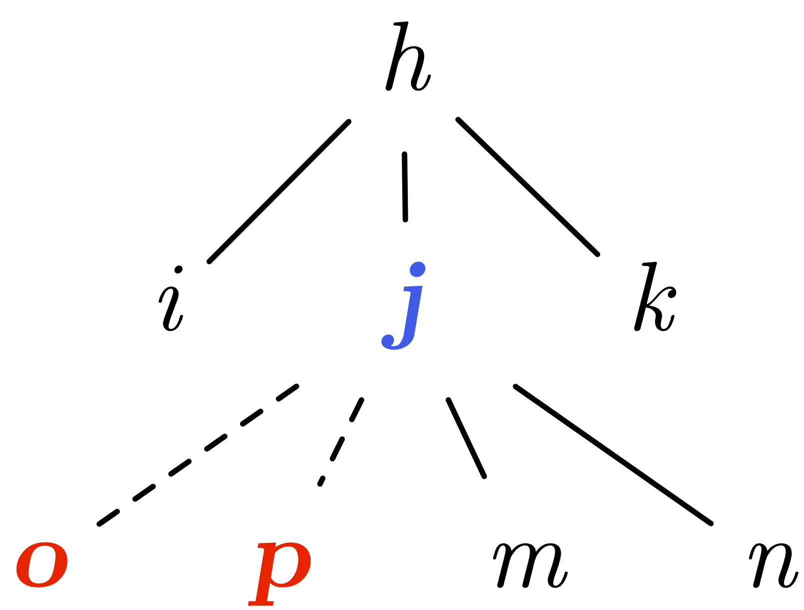
Tree Operations

Remove

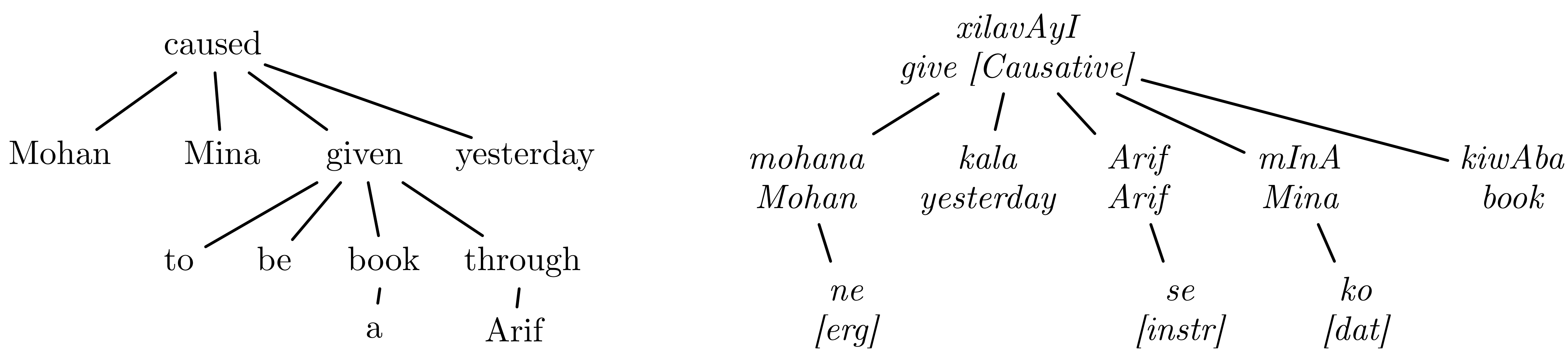


Tree Operations

Remove



Resolving Divergence

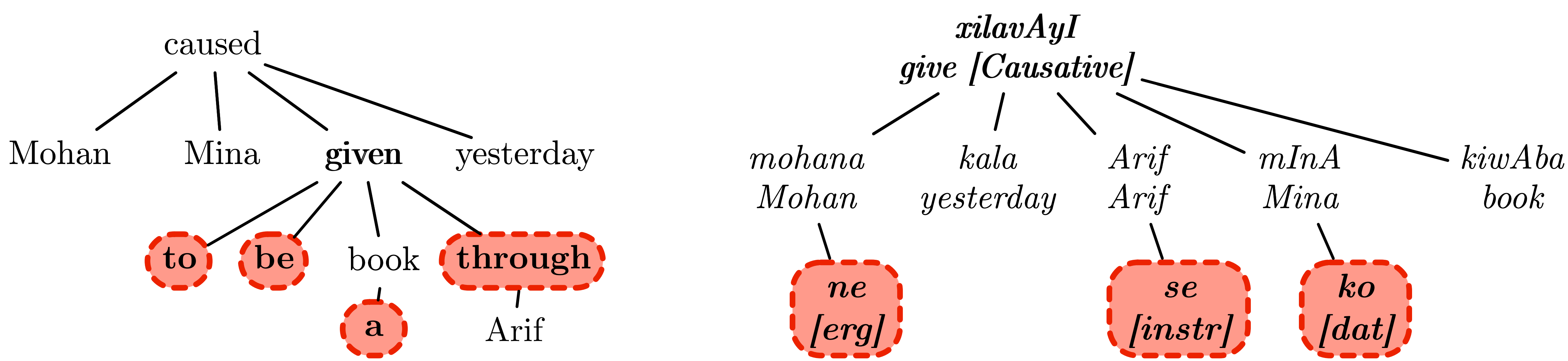


English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Resolving Divergence

Detect Spontaneous Nodes / Remove

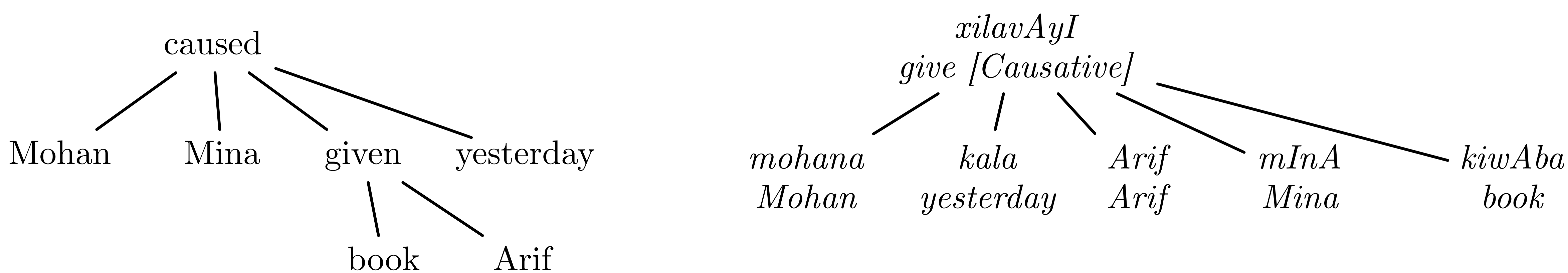


English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Resolving Divergence

Detect Spontaneous Nodes / Remove

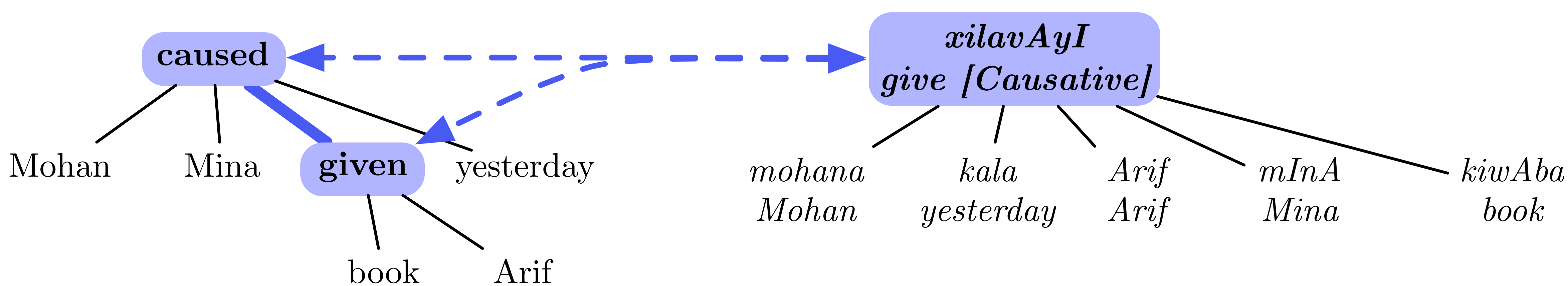


English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Resolving Divergence

Detect "Merge" Alignments / Merge

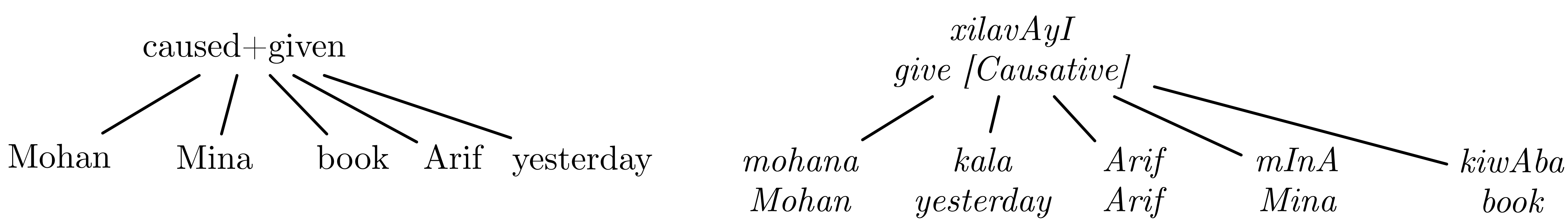


English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Resolving Divergence

Detect "Merge" Alignments / Merge

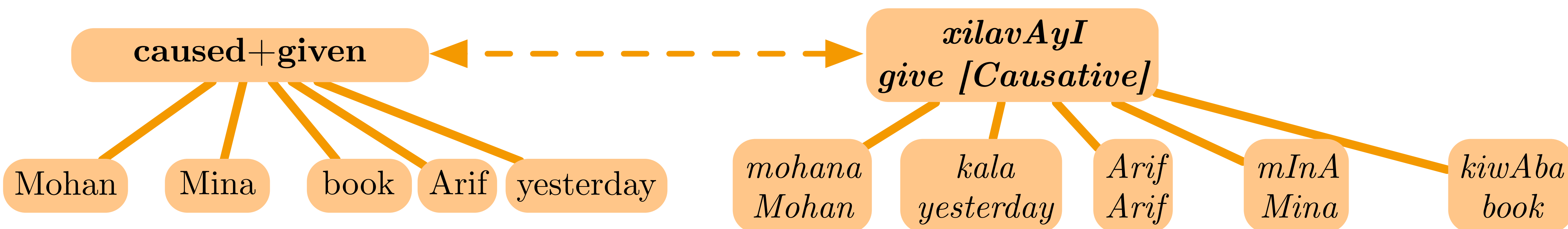


English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Resolving Divergence

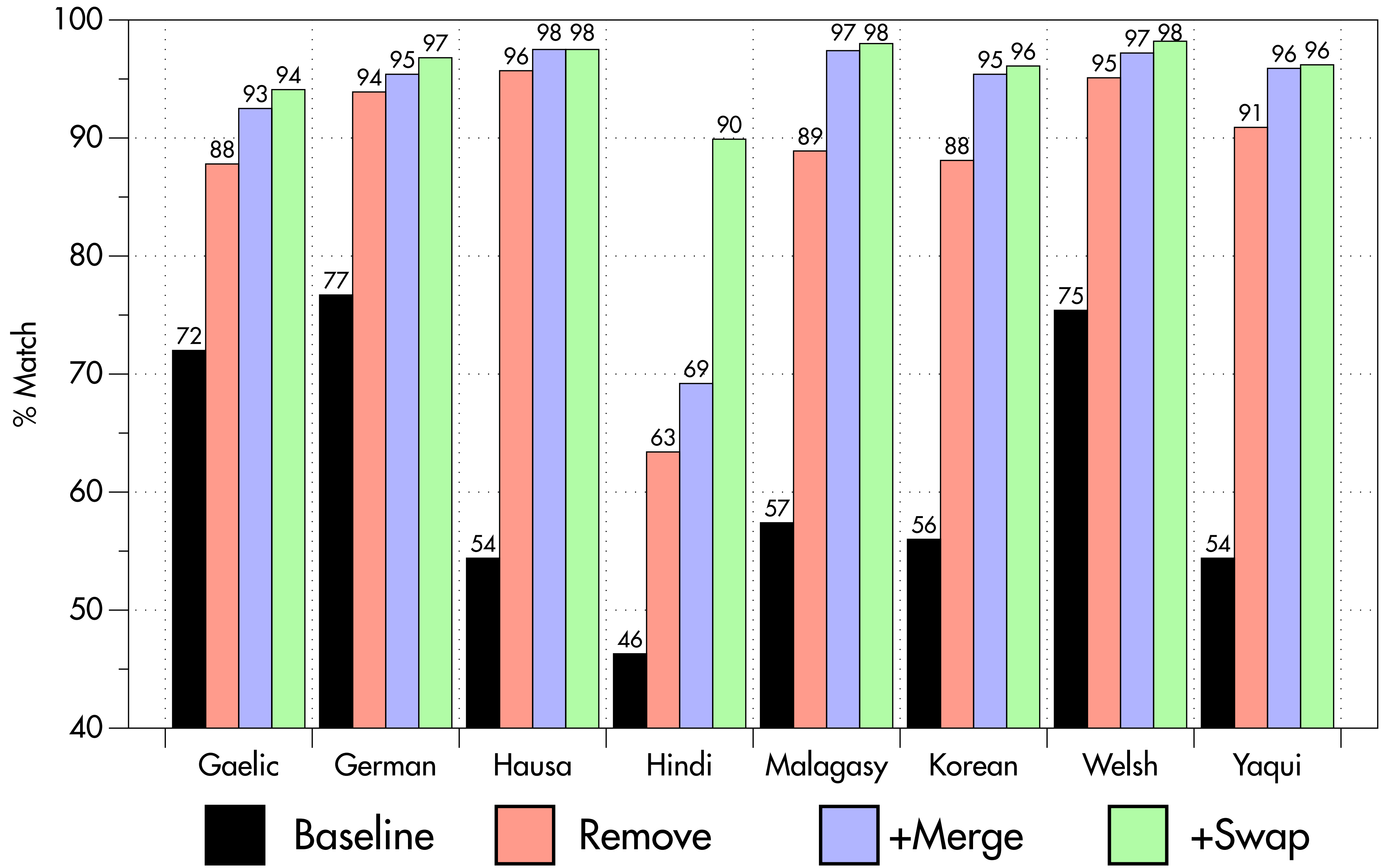
All Remaining Alignments Match



English:
Mohan caused Mina to be given a book through Arif yesterday

Hindi/Urdu:
mohana ne kala Arif se mInA ko kiwAba xilavAyI

Measuring Divergence



Learning Divergence

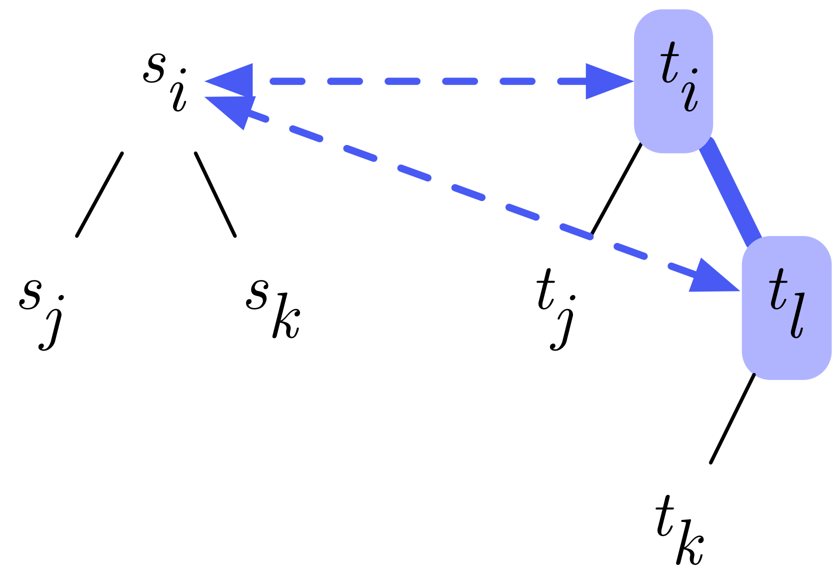
Learning Divergence

- Measured swaps, merges, and removals

Learning Divergence

- Measured swaps, merges, and removals
- Analyze the patterns of operations to learn post-processing rules

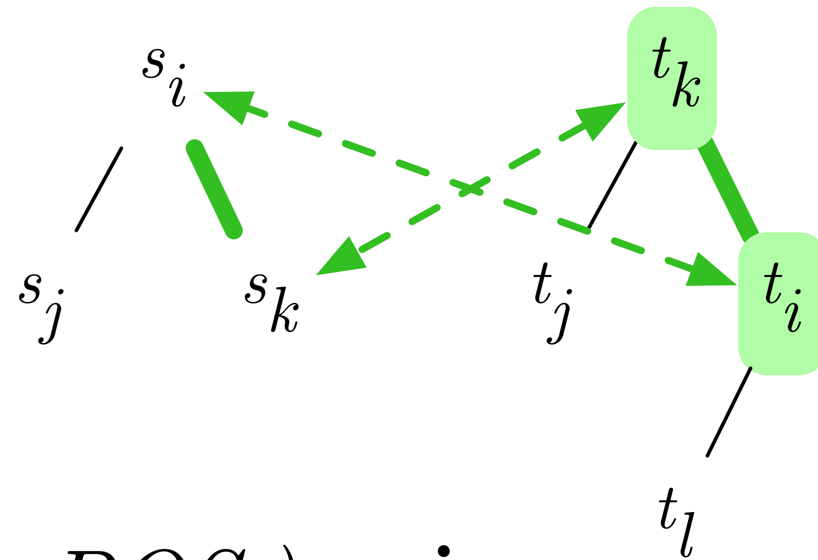
Multiply-Aligned Tokens



- For each POS_i, measure attachment direction
- At test time, choose head token from previous results

```
POSi → P(right) = 75%
POSi → P(left)  = 25%
```

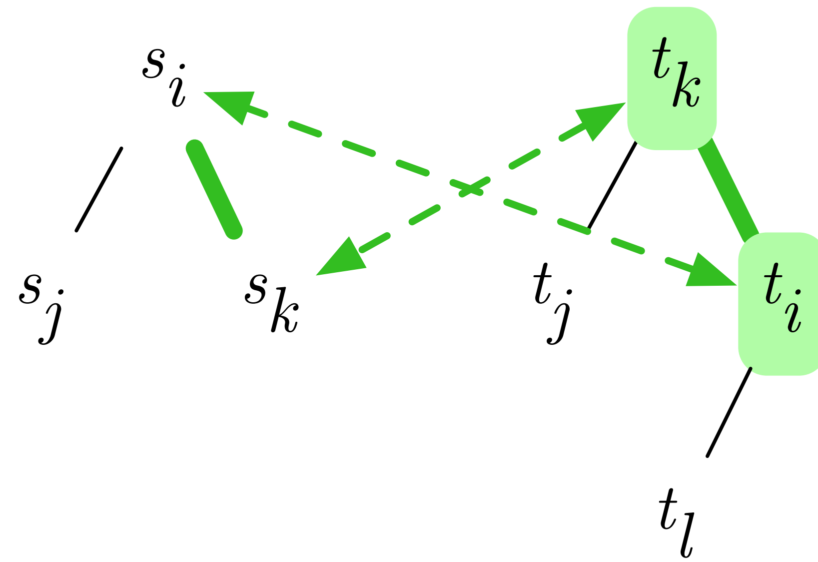
Swapped Tokens



- For each $(POS_i,\ POS_j)$ pair
  - Measure frequency of ***swap*** operation
- Apply swap at test time if
  - Occurs more than 3 times
  - More than 60% of occurrences:

```
(POSi,POSj) → P(swap) = 75%  [25/33]
(POSl,POSm) → P(swap) = 10%  [1/10]
(POSp,POSq) → P(swap) = 100% [1/1]
```

Spontaneous Tokens



- For each lexical item (t_i)
  - Measure attachment direction
- At test time:
  - Attach in majority direction
  - Backoff: attach in overall language-preferred direction

```
tn → P(right) = 75%
tm → P(left)  = [unseen]
Poverall(left) = 54%
```

Applying Rules to Projection

Applying Rules to Projection

- Two baselines:

Applying Rules to Projection

- Two baselines:
    - Prefer leftward attachment for merge/spontaneous

Applying Rules to Projection

- Two baselines:
  - Prefer leftward attachment for merge/spontaneous
  - Prefer rightward attachment for merge/spontaneous

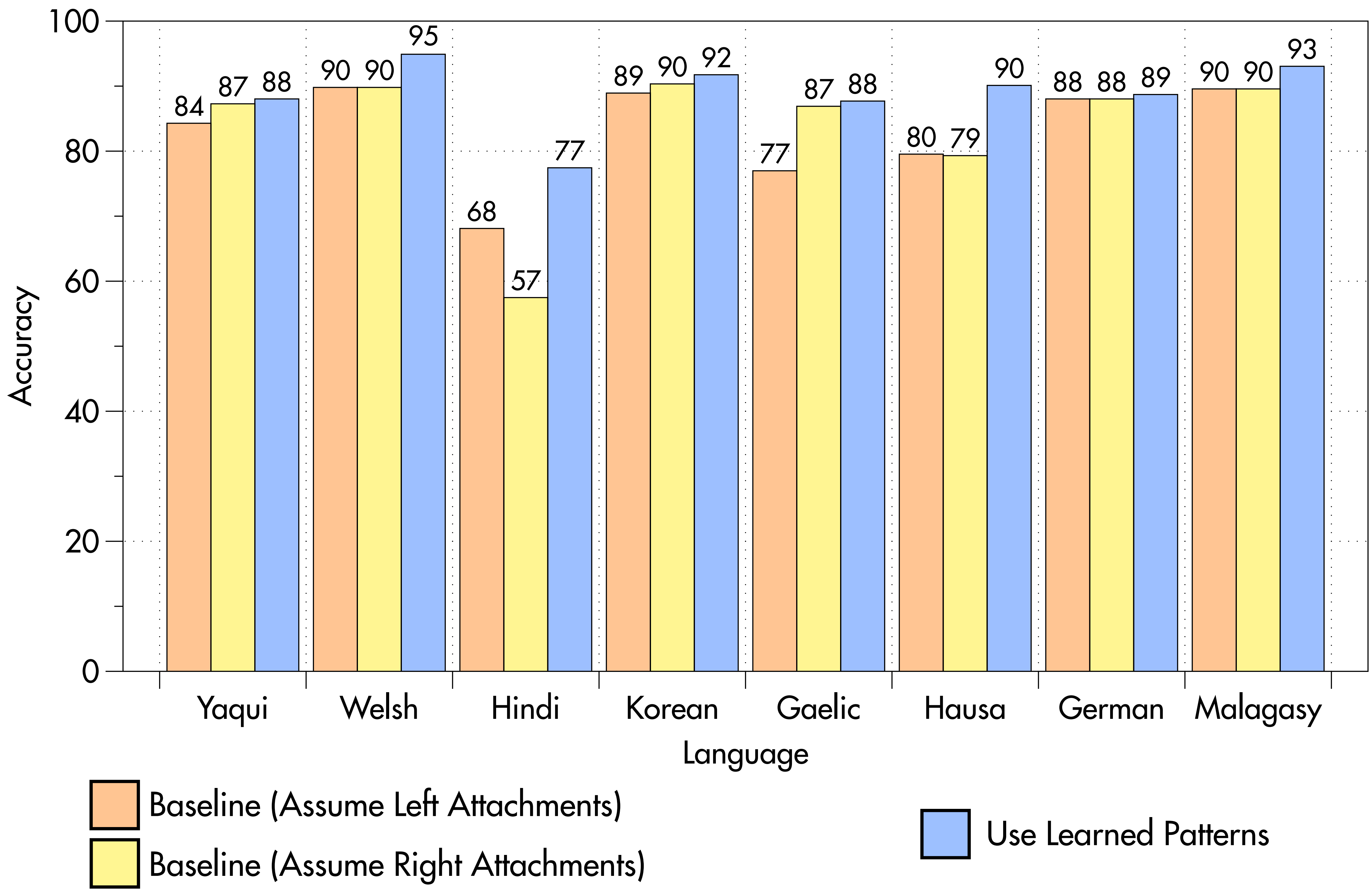
Applying Rules to Projection

- Two baselines:
  - Prefer leftward attachment for merge/spontaneous
  - Prefer rightward attachment for merge/spontaneous
  - No swap handling

Applying Rules to Projection

- Two baselines:
  - Prefer leftward attachment for merge/spontaneous
  - Prefer rightward attachment for merge/spontaneous
  - No swap handling
- Use learned merge, spontaneous, and swap rules

Rule-Enhanced Projection



(Re)Informing the Parser

Projection Options

Baseline (Assume Left Attachments)

Baseline (Assume Right Attachments)

Use Learned Patterns

(Re)Informing the Parser

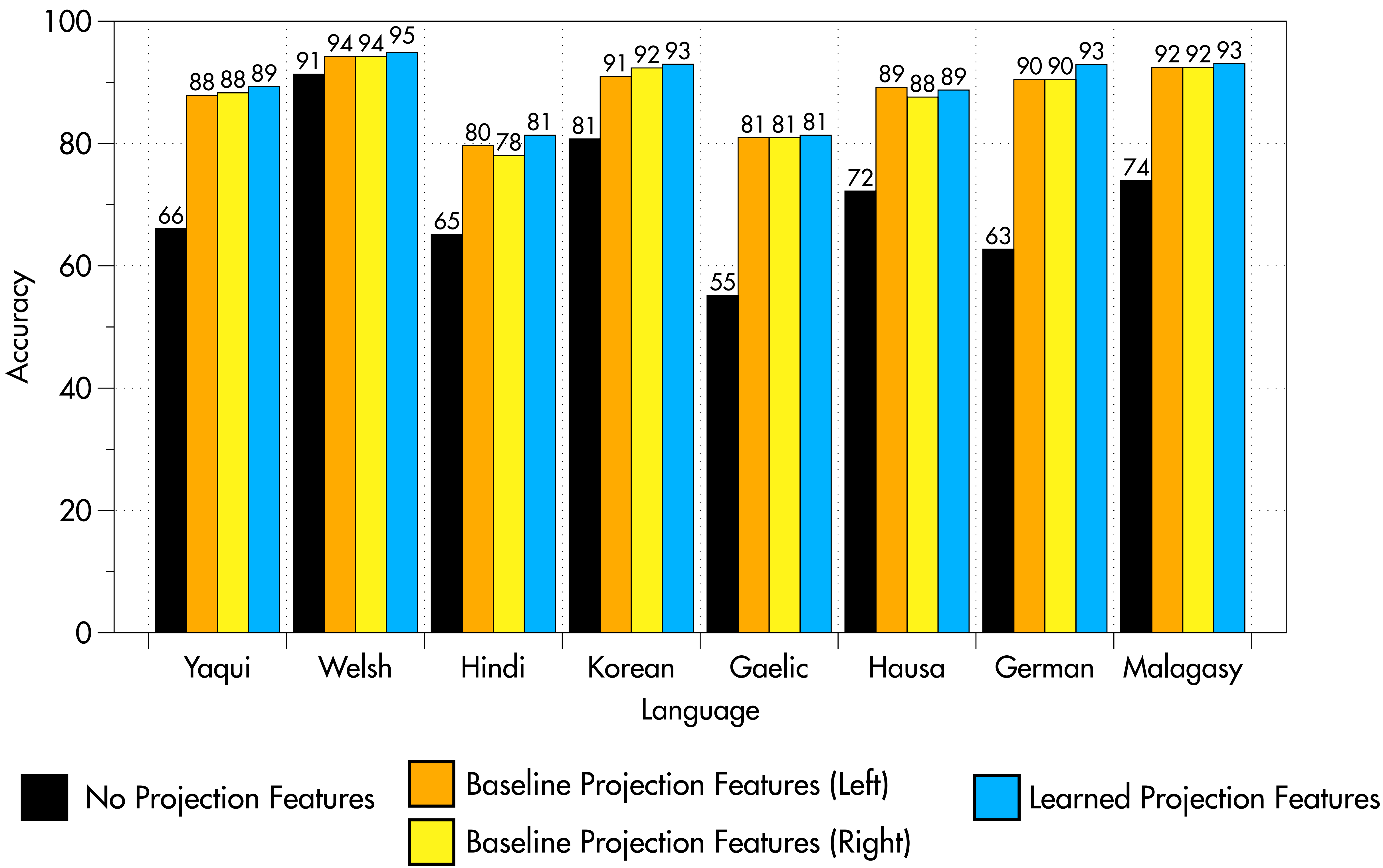
Projection Options

Baseline (Assume Left Attachments)
Baseline (Assume Right Attachments)
Use Learned Patterns

Parser Options

Baseline (Assume Left Attachments)
Baseline (Assume Right Attachments)
Use Learned Patterns
No Projection Features

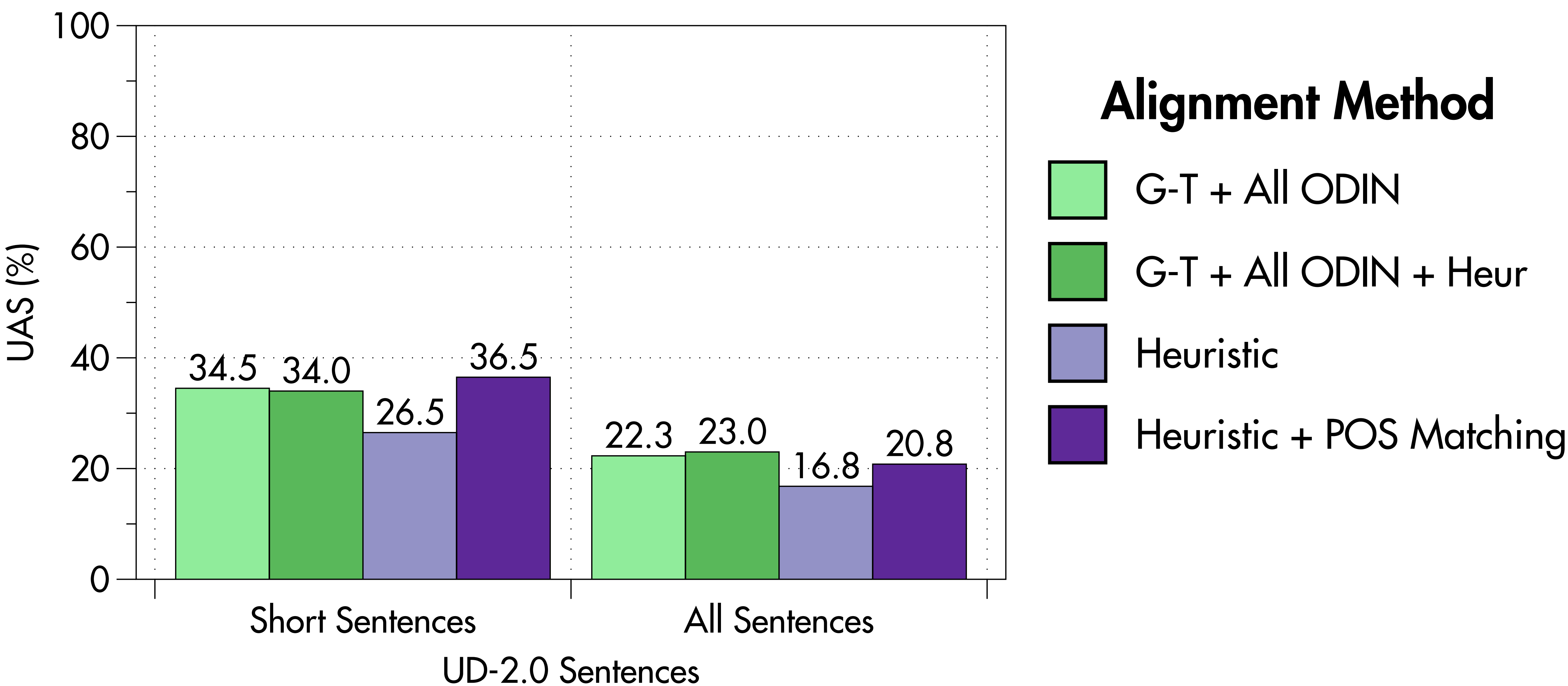
Parser with Improved Projections



Monolingual DS Parsing

- Use IGT-projected DSs to train monolingual parser
- Evaluate parser on the Universal Dependency corpus

Monolingual DS Parsing



Outline

- Previous Work
- Methodology
- Tasks
- Conclusion

Summary of Results

Word Alignment

- **0.85** F_1 score for heuristic alignment
- **0.83** F_1 score for improved statistical alignment
- **0.49** F_1 score for traditional approach

Summary of Results

Part-of-Speech Tagging

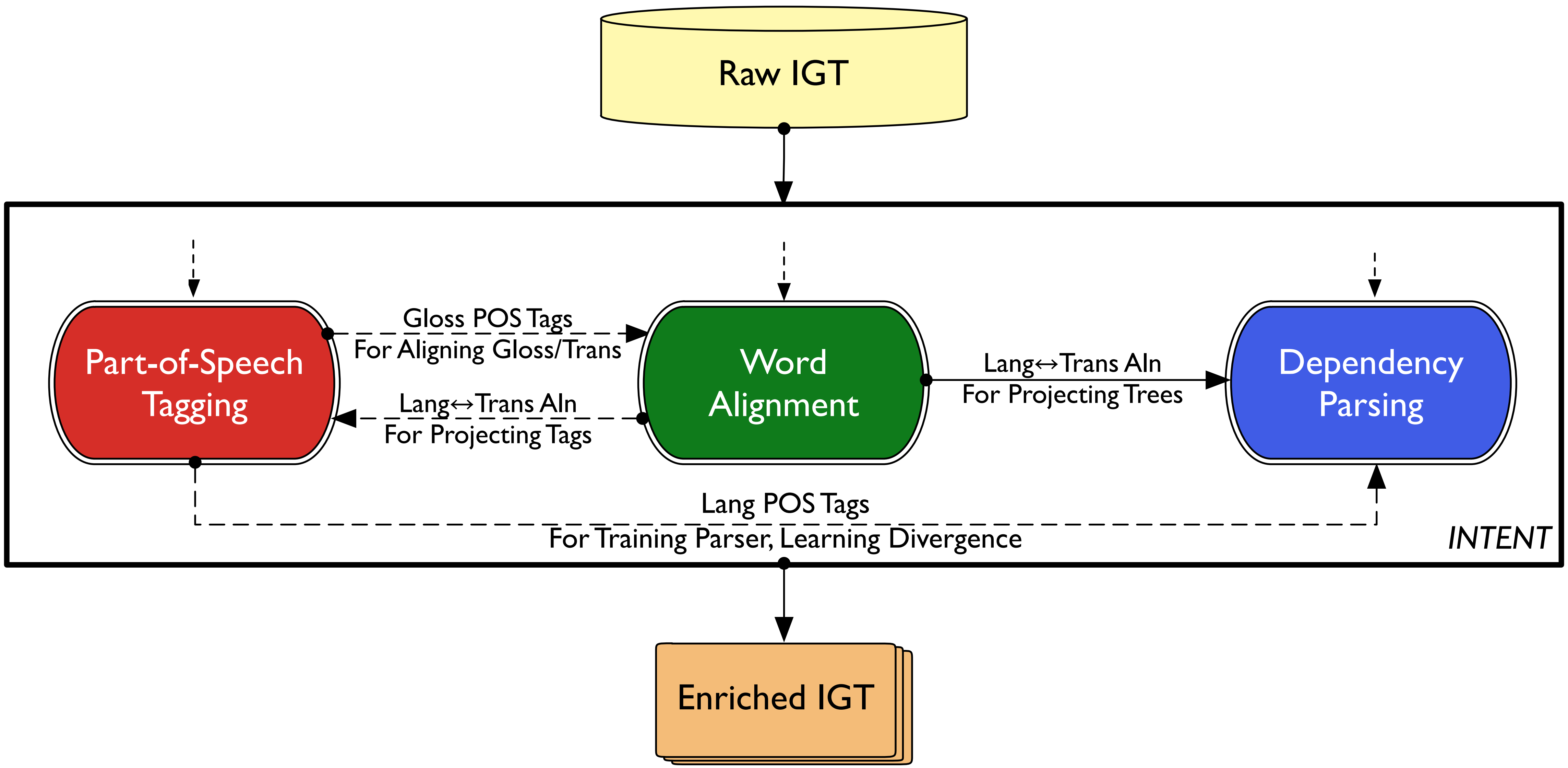
- **92%** accuracy on IGT
  - With classifier trained on manual gloss-line tags
  - **67%** using projection
- **70%** accuracy on monolingual data
  - Using classifier-bootstrapped taggers
  - **56%** using projection

Summary of Results

Dependency Parsing

- Analyzed language divergence
- **87%** accuracy for projection-feature enhanced parser
  - **84%** for projection alone
  - **67%** for baseline parser
- **89%** accuracy for enhanced parser w/rewrite rules
  - **88%** accuracy for enhanced projection

The INTENT System



Using INTENT

- Software package is available
- Code available at rgeorgi.co/intent
  - Online demo at rgeorgi.co/intentweb

Impact of INTENT

Impact of INTENT

- Used to enrich ODIN v2.1

Impact of INTENT

- Used to enrich ODIN v2.1
- Used at UW Linguistics Seminar SPR’15:
  - Computational Methods in Language Documentation

Impact of INTENT

- Used to enrich ODIN v2.1
- Used at UW Linguistics Seminar SPR'15:
  - Computational Methods in Language Documentation
- Being used to visualize enriched data in ODIN editor

Impact of INTENT

- Used to enrich ODIN v2.1
- Used at UW Linguistics Seminar SPR'15:
  - Computational Methods in Language Documentation
- Being used to visualize enriched data in ODIN editor
- Will be used for AGGREGATION Phase 2

Future Work

Future Work

- **Word Alignment:**
  - Use IGT-extracted alignments to bootstrap parallel data
  - “Clue-Based” alignment *(Tiedemann 2003)*

Future Work

- **Word Alignment:**
  - Use IGT-extracted alignments to bootstrap parallel data
  - "Clue-Based" alignment *(Tiedemann 2003)*

- **POS tagging:**
  - Use extracted POS tags to constrain induction approaches
    - *(Haghighi & Klein 2006, Mann & McCallum 2008)*

Future Work

- **Word Alignment:**
    - Use IGT-extracted alignments to bootstrap parallel data
    - "Clue-Based" alignment *(Tiedemann 2003)*

- **POS tagging:**
    - Use extracted POS tags to constrain induction approaches
        - *(Haghighi & Klein 2006, Mann & McCallum 2008)*

- **Dependency Parsing**
    - Use modified parser for partial trees *(Spreyer & Kuhn, 2009)*
    - Clustering/Similarity approaches *(Koo et. al, 2008; Mirroshandel et. al., 2012)*

Conclusion

Conclusion

- Utilized IGT's unique format to provide improvements over uninformed methods

Conclusion

- Utilized IGT's unique format to provide improvements over uninformed methods
- Created generalized IGT enrichment system covering 1,500+ languages

Conclusion

- Utilized IGT's unique format to provide improvements over uninformed methods
- Created generalized IGT enrichment system covering 1,500+ languages
- Demonstrated potential for IGT-bootstrapped NLP tools in resource-poor settings

Thank You

Related Publications

Xia, F., &al. — **Enriching a massively multilingual database of interlinear glossed text.** *Language Resources and Evaluation* (2016).

Xia, F., &al. — **Enriching, Editing, and Representing Interlinear Glossed Text.** CICLing 2015.

Georgi, R., &al. — **Enriching Interlinear Text using Automatically Constructed Annotators.** LaTeCH 2015.

Georgi, R., &al.— **Capturing divergence in dependency trees to improve syntactic projection.** *Language Resources and Evaluation* (2014).

Georgi, R., &al. — **Improving Dependency Parsing with Interlinear Glossed Text and Syntactic Projection.** COLING 2012.

Georgi, R., &al. — **Measuring the Divergence of Dependency Structures Cross-Linguistically to Improve Syntactic Projection Algorithms**. LREC 2012.

Georgi, R., &al. — **Enhanced and Portable Dependency Projection Algorithms Using Interlinear Glossed Text.** ACL 2013.

Related Software

INTENT	rgeorgi.co/intent
ODIN Editor	rgeorgi.co/xigtedit
ODIN v2.1	rgeorgi.co/odin

Key References

1. Nicoletta Calzolari, Claudia Soria, Irene Russo, Francesco Rubino, and Riccardo Del Gratta. 2010. The Language Resources and Evaluation (LRE) Map. flarenet.eu.

2. Paul M Lewis, Gary F Simons, and Charles D Fennig, editors. 2016. Ethnologue. SIL International, Dallas, Texas, 19 edition. https://www.ethnologue.com/

3. Fei Xia, William D Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation*:1–29, January.

4. Jörg Tiedemann. 2003. Combining clues for word alignment. In *The Tenth Conference of The European Chapter of the Association for Computational Linguistics*, volume 1, pages 339–346, Morristown, NJ, USA, April. Association for Computational Linguistics.

5. William D Lewis and Fei Xia. 2010. Developing ODIN: A Multilingual RePOSitory of Annotated Language Data for Hundreds of the World's Languages. Literary and Linguistic Computing, 25(3):303–319, August.

Key References

6. Emily M Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Propertie. In *Proceedings of the ACL workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities*, Sofia, Bulgaria.

7. Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Sofia, Bulgaria.

8. Bonnie Jean Dorr. 1994. Machine translation divergences: a formal description and proposed solution. Computational Linguistics, 20:597–633, December.

9. Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. Natural Language Engineering, 11(03):311–325, September.

Key References

10. Seyed Abolghasem Mirroshandel, Alexis Nasr, and Joseph Le Roux. 2012. Semi-supervised dependency parsing using lexical affinities. In ACL 2012, pages 777–785.

IGT References

1. Matthew S Dryer. 2007. Noun phrase structure. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, volume 2, pages 151–205. Language typology and syntactic description, Cambridge, United Kingdom, 2nd edition, October.

2. Dorothee A Beermann and Lars Hellan. 2002. VP-Chaining in Oriya. In Stanford Linguistic Association and CSLI.

3. Pierre Lafitte. 1962. Grammaire basque. Editions des "amis du musee basque" et “Ikas,” Bayonne, edition.

POS Projection Confusion Matrix

	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	PREC
ADJ	**57**	1	0	0	2	4	0	6	0	0	0	0.81
ADP	0	**52**	2	10	2	0	0	6	2	2	0	0.68
ADV	0	2	**69**	0	2	5	0	0	0	0	0	0.88
CONJ	0	0	0	**20**	0	0	0	0	0	0	0	1
DET	2	6	0	2	**370**	0	0	6	2	0	0	0.95
NOUN	4	1	10	0	2	**649**	2	4	0	26	0	0.93
NUM	0	0	0	0	0	0	**16**	2	0	0	0	0.89
PRON	0	0	2	0	14	0	0	**219**	0	6	0	0.91
PRT	0	4	0	0	0	0	0	0	**26**	2	0	0.81
VERB	1	2	1	0	0	20	0	2	0	**574**	0	0.96
X	0	0	0	0	0	0	0	25	0	0	**0**	0
Unaligned	*8*	*48*	*24*	*4*	*56*	*58*	*2*	*50*	*18*	*114*	*0*	
% Unaligned	*11.1*	*41.4*	*22.2*	*11.1*	*12.5*	*7.9*	*10*	*15.6*	*37.5*	*15.7*	*0*	
REC	0.79	0.45	0.64	0.56	0.83	0.88	0.8	0.68	0.54	0.79	0	

Classifier Confusion Matrix

	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	PREC
ADJ	**18**	0	0	0	0	0	0	2	0	0	0	0.9
ADP	0	**40**	1	6	2	0	0	2	9	1	0	0.66
ADV	1	0	**27**	0	1	0	0	0	0	0	0	0.93
CONJ	0	0	0	**4**	0	0	0	0	0	0	0	1
DET	0	0	0	0	**112**	3	0	1	1	0	0	0.96
NOUN	3	0	6	0	3	**204**	1	2	1	7	0	0.9
NUM	0	0	0	0	0	0	**5**	0	0	0	0	1
PRON	0	0	0	0	3	0	0	**93**	0	0	0	0.97
PRT	0	0	0	0	0	0	0	0	**3**	0	0	1
VERB	0	1	0	0	0	4	0	0	0	**211**	1	0.97
X	0	0	0	0	0	0	0	0	0	0	0	0
REC	0.82	0.98	0.79	0.4	0.92	0.97	0.83	0.93	0.21	0.96	0	