Evaluation of Finite State Morphological Analyzers Based on Paradigm Extraction from Wiktionary

Ling Liu
Mans Hulden

FSMNLP
Sep 6, 2017

Outline

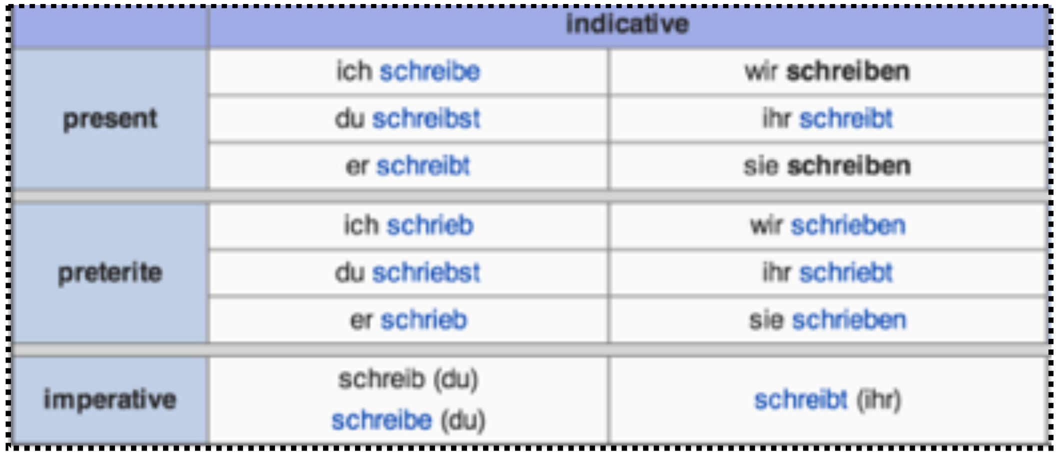
- Motivation of the Study
- How the Morphological Analyzer works
- Data
- Evaluation and Result

Outline

- Motivation of the Study
- How the Morphological Analyzer works
- Data
- Evaluation and Result

Motivation of the Study

- Wiktionary: morphological inflection tables for many languages

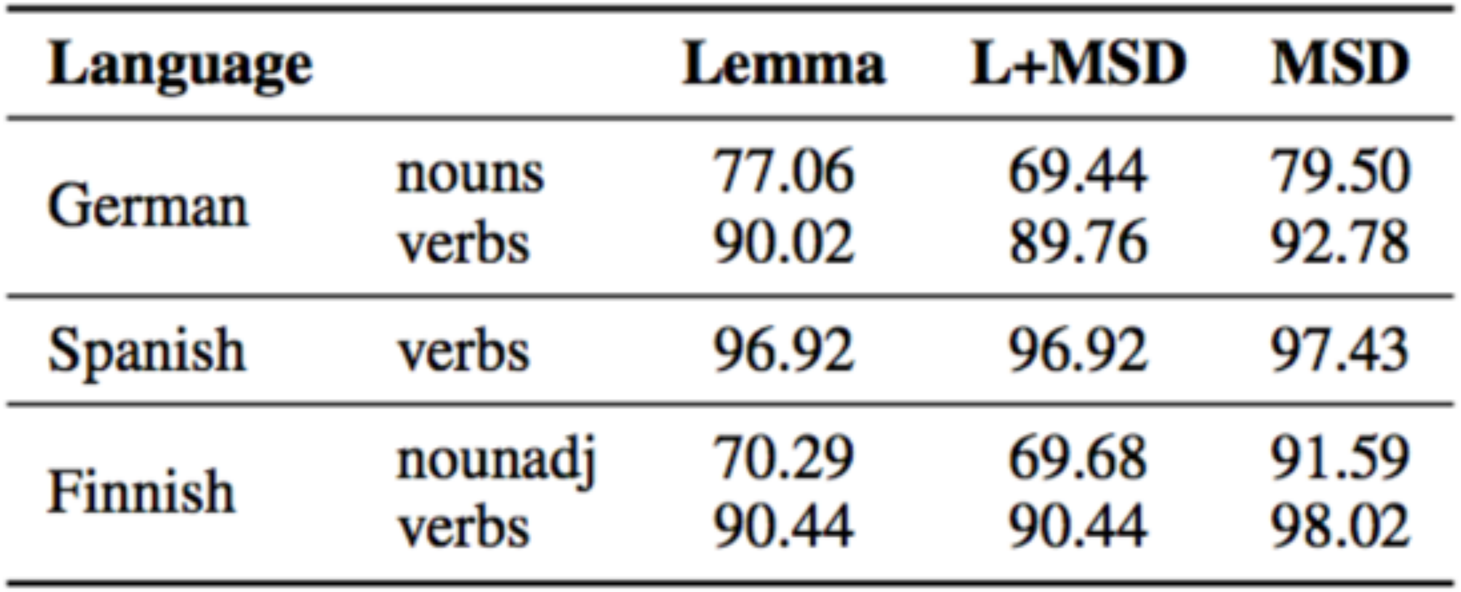


	indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)

- Wiktionary Morphological Database: 350 languages

Motivation of the Study

- Forsberg and Hulden (2016): a method to convert morphological inflection tables into unweighted and weighted finite transducers for parsing and generation
    - Evaluated on German, Spanish, Finnish



Language		Lemma	L+MSD	MSD
German	nouns	77.06	69.44	79.50
	verbs	90.02	89.76	92.78
Spanish	verbs	96.92	96.92	97.43
Finnish	nounadj	70.29	69.68	91.59
	verbs	90.44	90.44	98.02

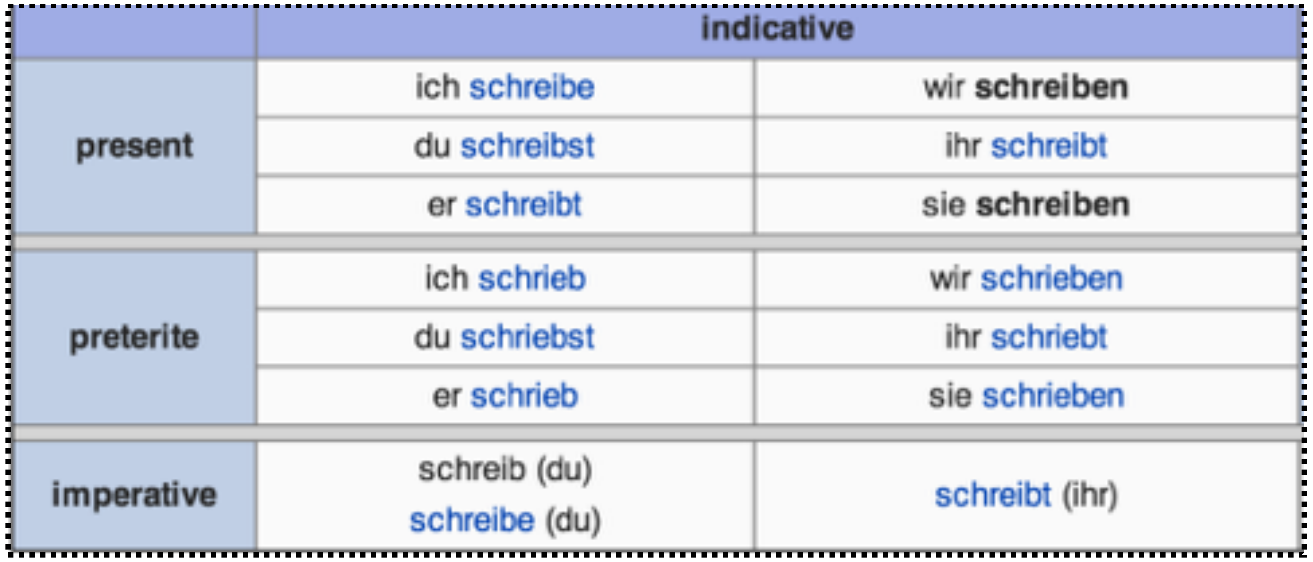
Evaluation of the weighted model (Forsberg and Hulden, 2016)

Outline

- Motivation of the Study
- How the Morphological Analyzer works
- Data
- Evaluation and Result

Workflow of the Morphological Analyzer

Inflection examples



	indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)

Step 1

Paradigmatic Models

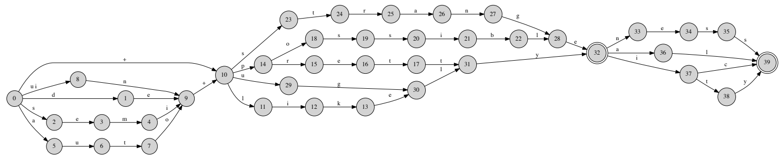
x_1 **i** x_2

x_1 **a** x_2

x_1 **u** x_2

Step 2

FST



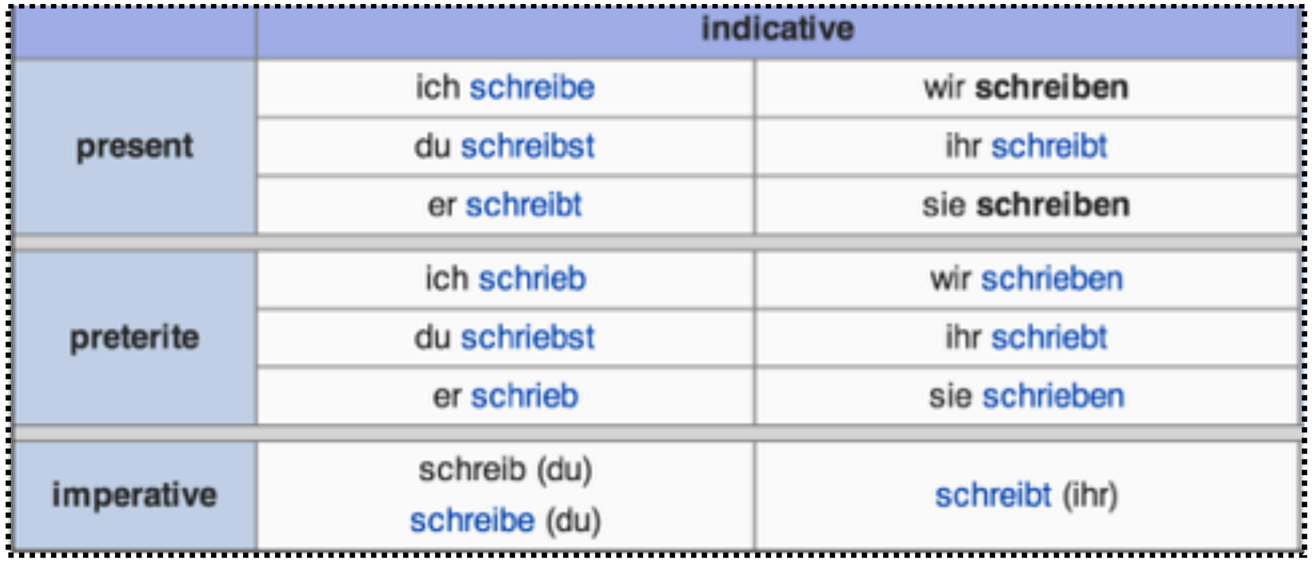
Step 3

Ranking Analyses

blargashed
1. blargash[V;PST]
2. blargash[V;V.PTCP;PST]
3. blargashe[V;PST]
…

Generalization from inflection tables

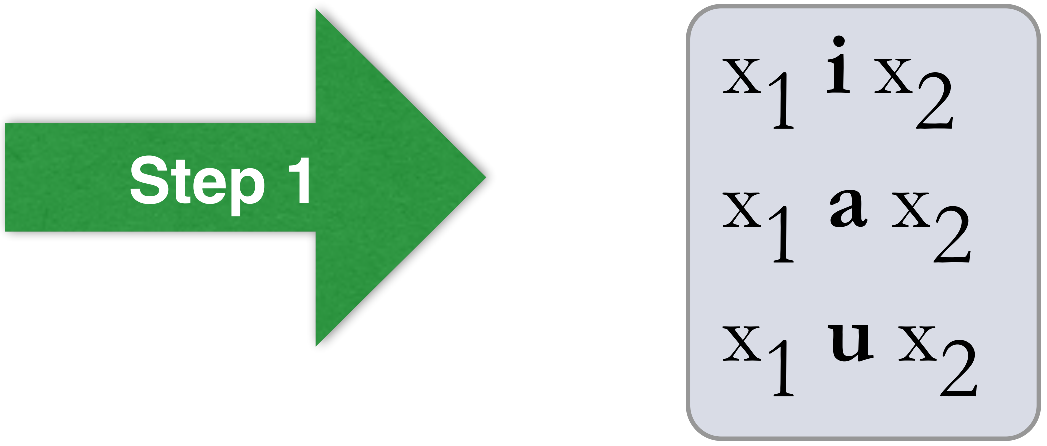
Inflection examples



	indicative	
present	ich schreibe	wir **schreiben**
	du schreibst	ihr schreibt
	er schreibt	sie **schreiben**
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)

Wiktionary
the free dictionary

Paradigmatic Models



$x_1\,\mathbf{i}\,x_2$

$x_1\,\mathbf{a}\,x_2$

$x_1\,\mathbf{u}\,x_2$

Generalization

The **common parts (stem)** are calculated by extracting the **Longest Common Subsequence** from related forms

inflection table

ring
rang
rung
rings
ringing

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

The **common parts (stem)** are calculated by extracting the **Longest Common Subsequence** from related forms

inflection table

r i ng
r a ng
r u ng
r i ng s
r i ng ing

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

The **common parts (stem)** are calculated by extracting the **Longest Common Subsequence** from related forms

inflection table

r i ng
r a ng
r u ng
r i ng s
r i ng ing

LCS = rng

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

The **common parts (stem)** are calculated by extracting the **Longest Common Subsequence** from related forms

inflection table

r i ng

r a ng

r u ng

r i ng s

r i ng ing

LCS = rng

$x_1 = \mathbf{r}$

$x_2 = \mathbf{ng}$

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table

r i ng
r a ng
r u ng
r i ng s
r i ng ing

$\underline{\quad}$ $\underline{\quad}$
x_1 x_2

LCS = **rng**

x_1 = **r**
x_2 = **ng**

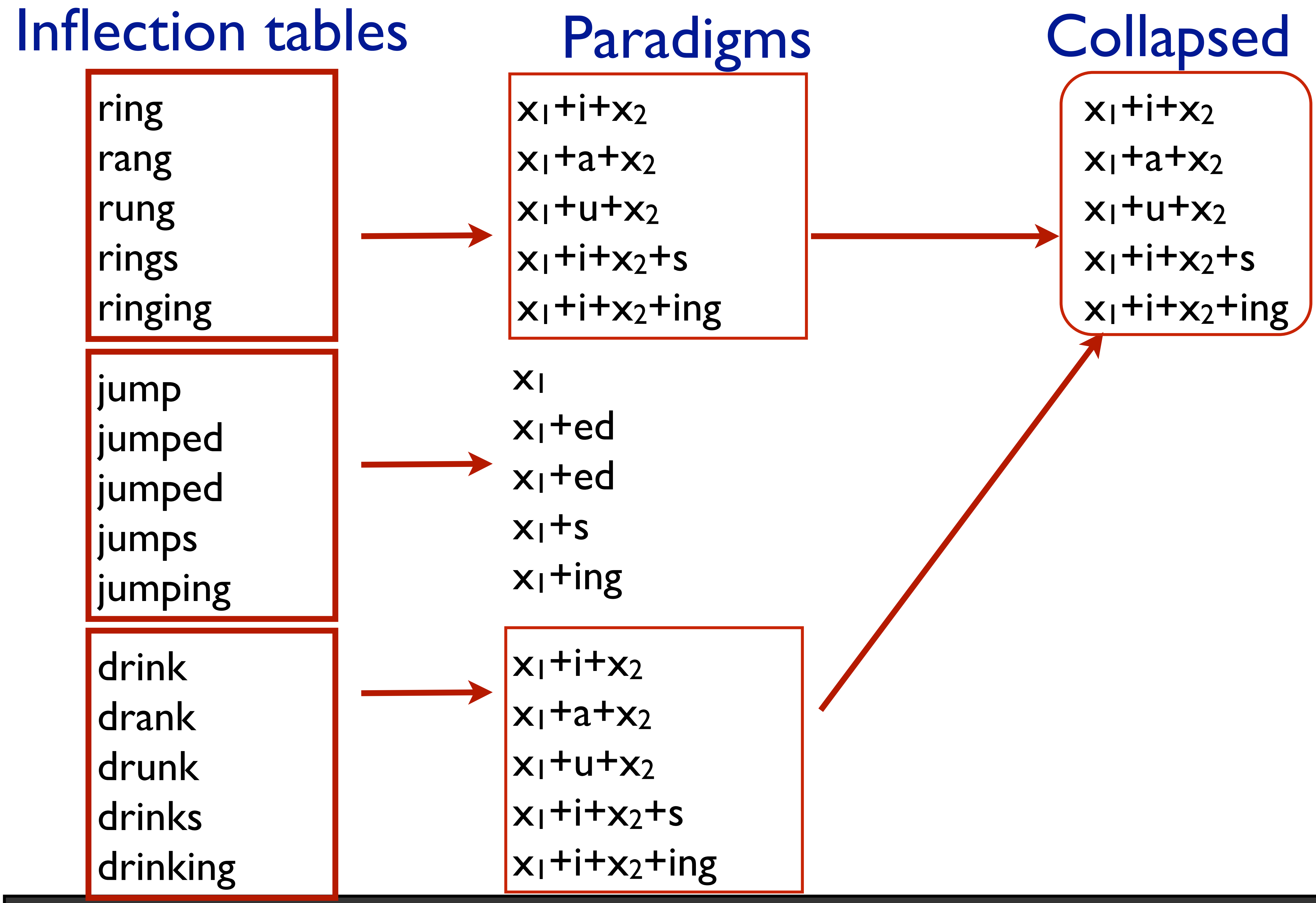
→

"paradigm"

$x_1 + i + x_2$
$x_1 + a + x_2$
$x_1 + u + x_2$
$x_1 + i + x_2 + s$
$x_1 + i + x_2 + ing$

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization



Inflection tables

ring
rang
rung
rings
ringing

jump
jumped
jumped
jumps
jumping

drink
drank
drunk
drinks
drinking

Paradigms

x_1+i+x_2
x_1+a+x_2
x_1+u+x_2
x_1+i+x_2+s
x_1+i+x_2+ing

x_1
x_1+ed
x_1+ed
x_1+s
x_1+ing

x_1+i+x_2
x_1+a+x_2
x_1+u+x_2
x_1+i+x_2+s
x_1+i+x_2+ing

Collapsed

x_1+i+x_2
x_1+a+x_2
x_1+u+x_2
x_1+i+x_2+s
x_1+i+x_2+ing

From paradigm to FST

Paradigmatic Models

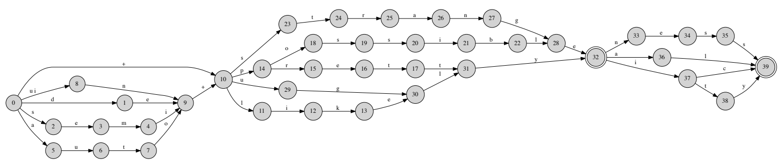
$x_1 \, \mathbf{i} \, x_2$

$x_1 \, \mathbf{a} \, x_2$

$x_1 \, \mathbf{u} \, x_2$

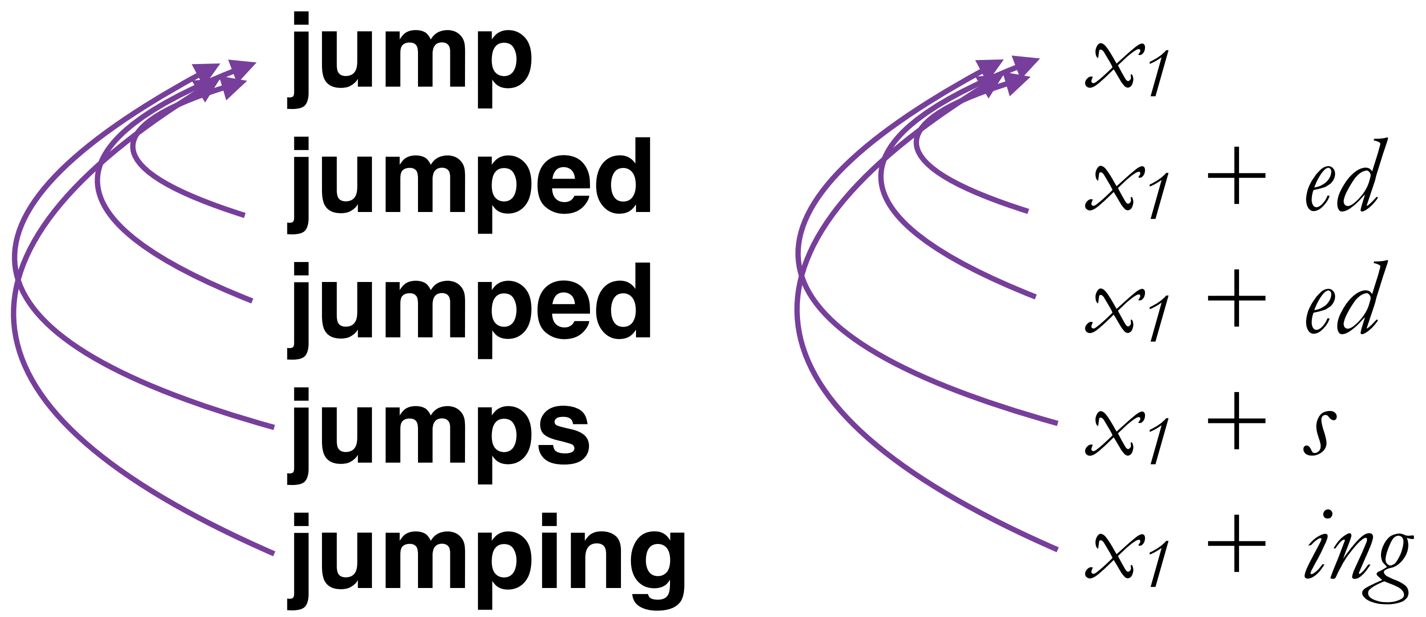
Step 2

FST



From paradigm to FST

Lemmatization

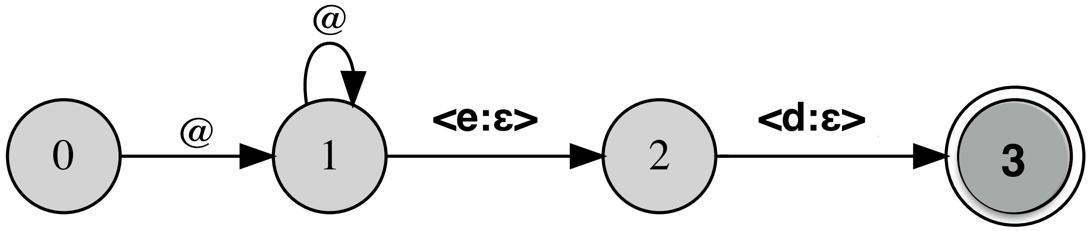


jump	x_1	**infinitive**
jumped	$x_1 + \mathit{ed}$	**simp past**
jumped	$x_1 + \mathit{ed}$	**past part**
jumps	$x_1 + \mathit{s}$	**simp pres 3sg**
jumping	$x_1 + \mathit{ing}$	**pre part**

From paradigm to FST

Lemmatization

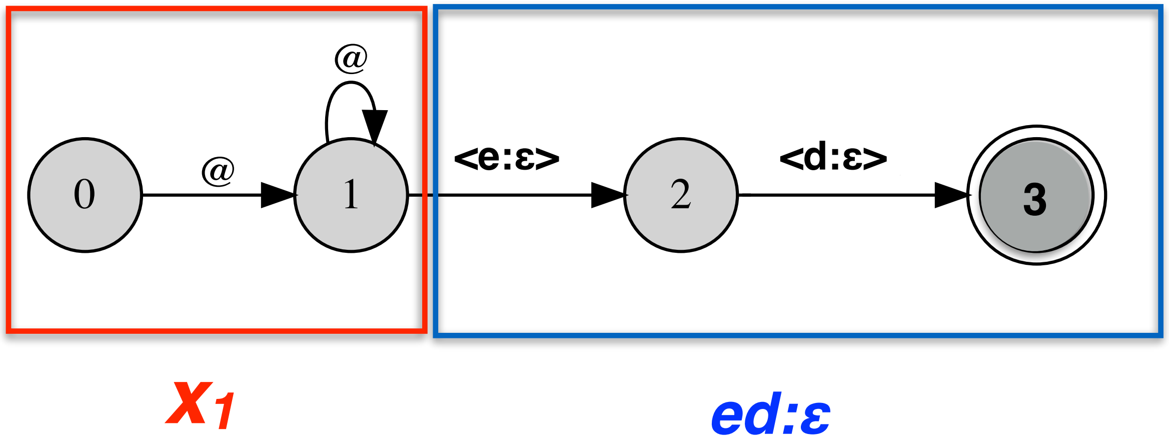
jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**



From paradigm to FST

Lemmatization

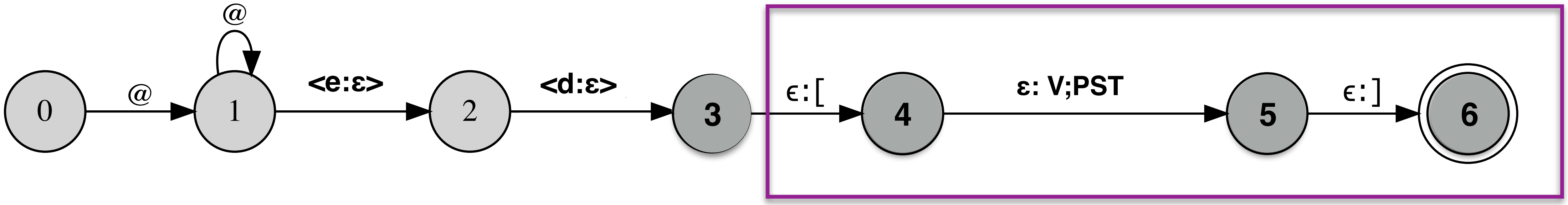
jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**



From paradigm to FST

Add inflection information

jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**

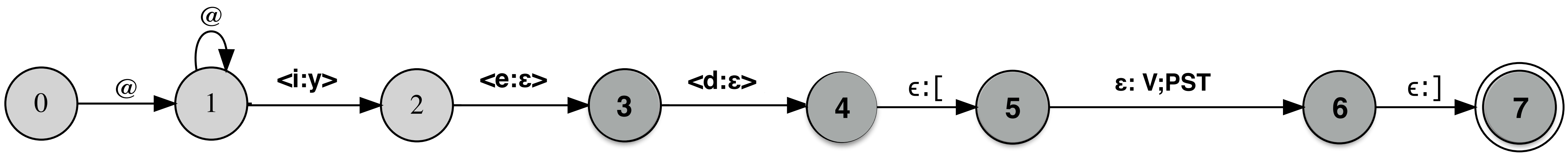


jumped > jump[V;PST]

From paradigm to FST

More lemmatization and analysis example

rely	$x_1 + y$	**infinitive**
relied	$x_1 + ied$	**simp past**
relied	$x_1 + ied$	**past part**
relies	$x_1 + ies$	**simp pres 3sg**
relying	$x_1 + ying$	**pre part**



relied > rely[V;PST]

Building the analyzer

Paradigm

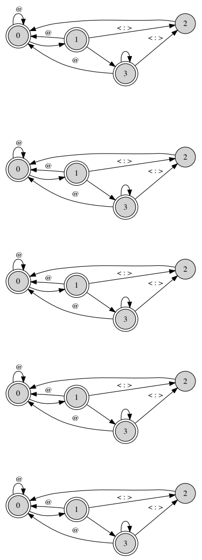
x_1+i+x_2

x_1+a+x_2

x_1+u+x_2

x_1+i+x_2+s

x_1+i+x_2+ing



analyzers

Paradigm

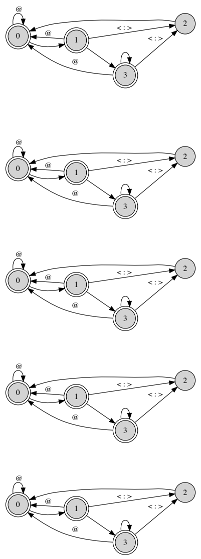
x_1

x_1+ed

x_1+ed

x_1+s

x_1+ing

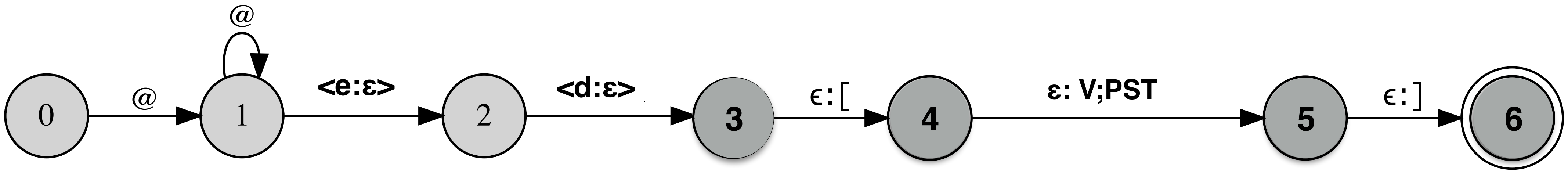


analyzers

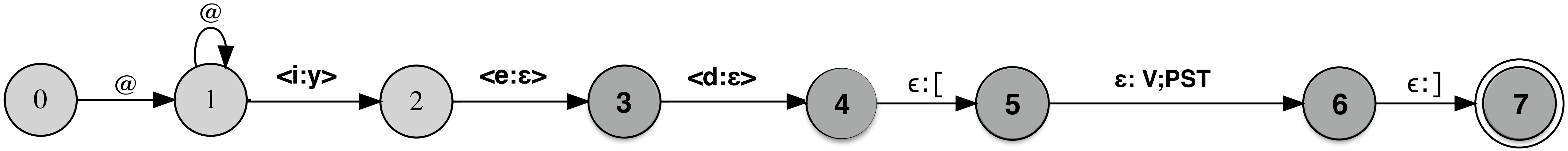
m transducers

Analyzer = $f_1 \cup f_2 \cup \ldots \cup f_1 \cup \ldots \cup f_m$

From paradigm to FST



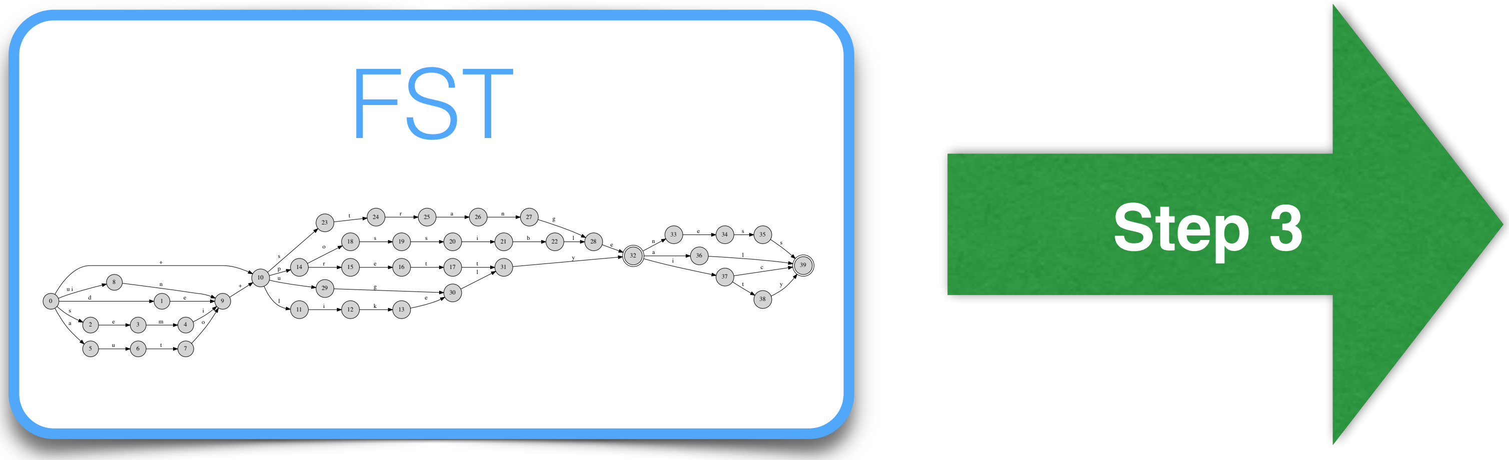
jumped > jump[V;PST]



relied > rely[V;PST]

verified	tried	died
verify[V;PST] verifi[V;PST] …	try[V;PST] tri[V;PST] …	dy[V;PST] di[V;PST] die[V;PST] …

Ranking Analyses



Ranking Analyses

blargashed

1. blargash[V;PST]
2. blargash[V;V.PTCP;PST]
3. blargashe[V;PST]

…

Language models over variables (WFSTs)

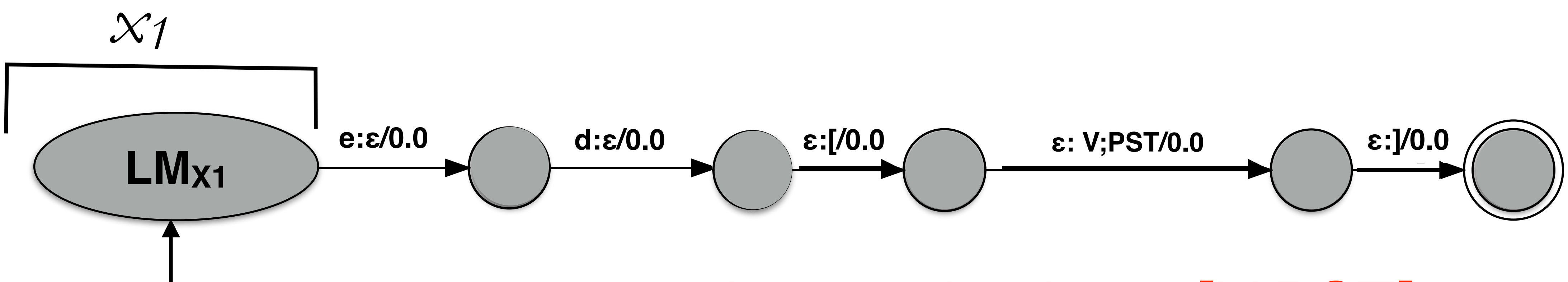
jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**

jump
watch
look
listen
work
ask
…

Infer a language model!

From paradigm to WFST

jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**

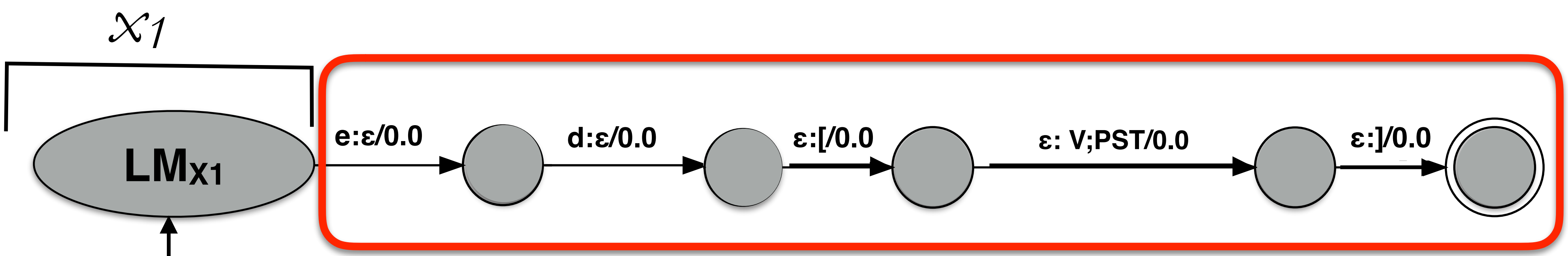


n-gram model

jumped > jump[V;PST]

From paradigm to WFST

jump	x_1	**infinitive**
jumped	$x_1 + ed$	**simp past**
jumped	$x_1 + ed$	**past part**
jumps	$x_1 + s$	**simp pres 3sg**
jumping	$x_1 + ing$	**pre part**



n-gram model

jumped > jump[V;PST]

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Example analysis (weighted)

Both are correct

rank	log_prob	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

verified

Outline

- Motivation of the Study
- How the Morphological Analyzer works
- Data
- Evaluation and Result

Data

- Wiktionary Morphological Database
- UniMorph project (https://unimorph.github.io/index.html)

- 55 Languages

- 19 Language groups

- 10 scripts

Outline

- Motivation of the Study
- How the Morphological Analyzer works
- Data
- Evaluation and Result

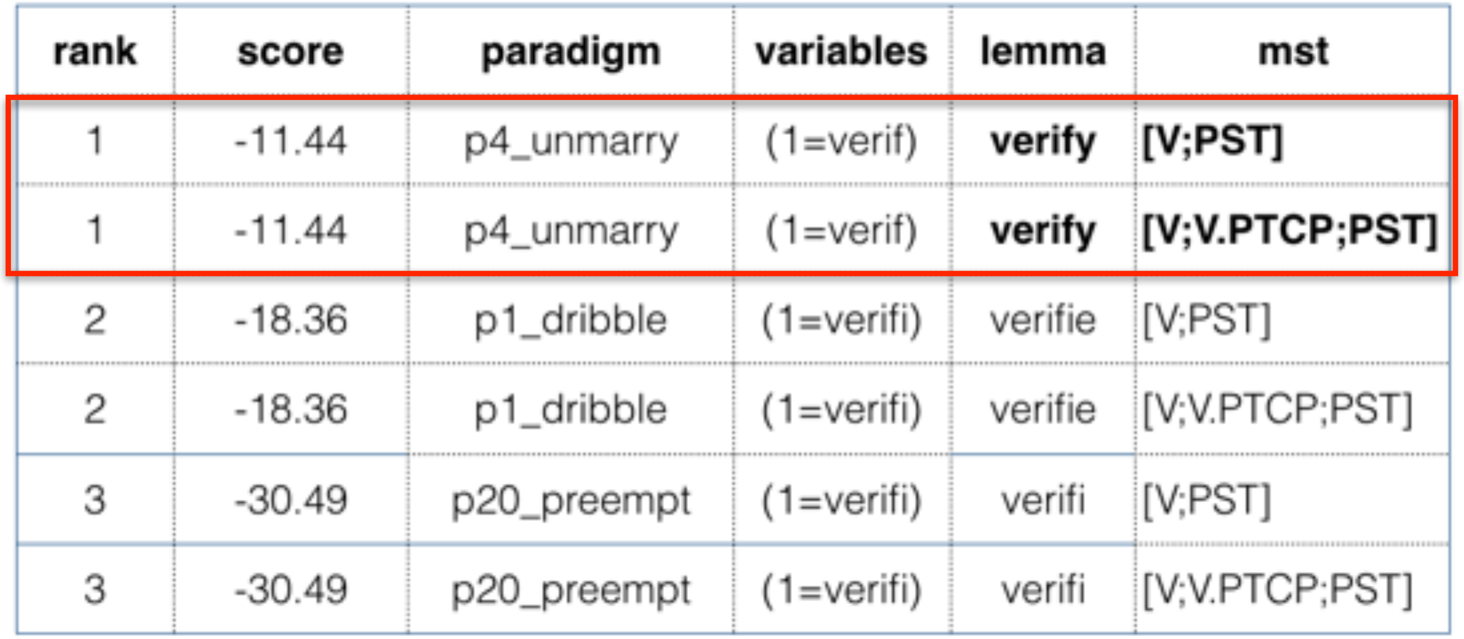
Evaluation Task

- Lemmatization and morphosyntactic information tagging

- 90% for training; and 10% for test (unless less than 50 inflection tables)
- The evaluation data is **disjoint** from the training data

- The first-ranked analyses

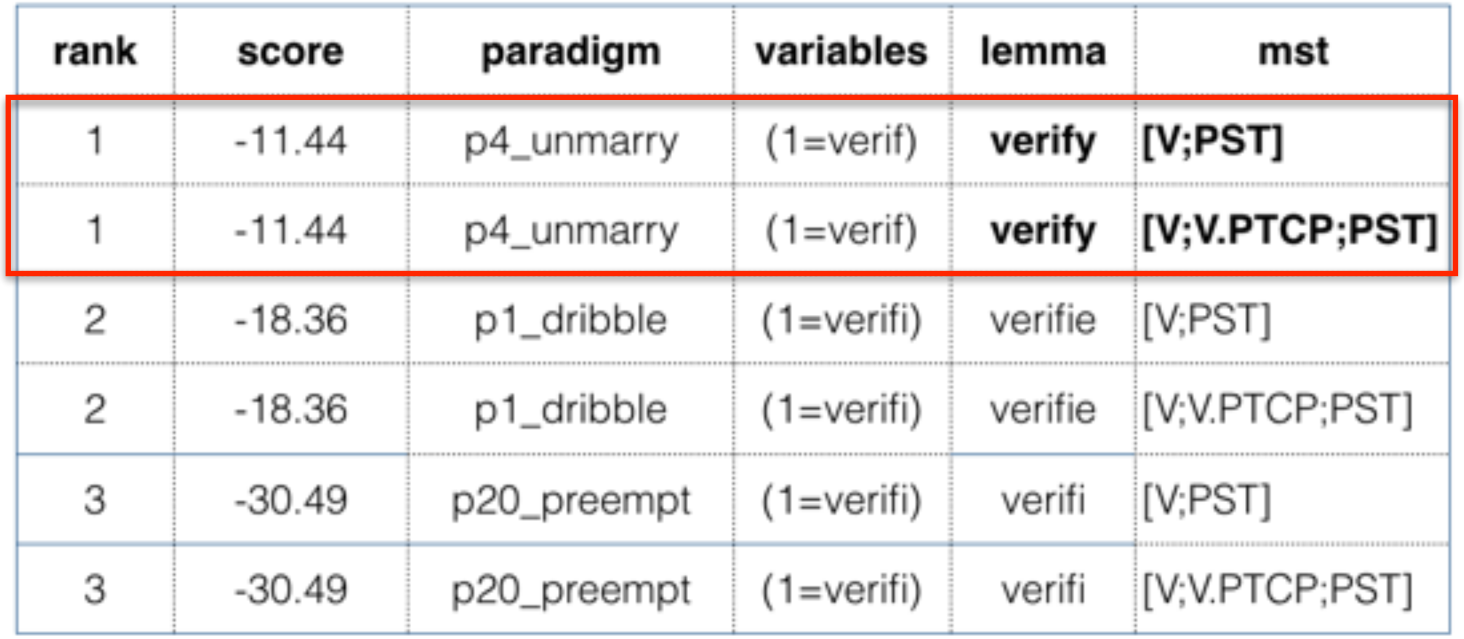
- Recall
  - lemma
  - lemma + POS
  - lemma + MST



rank	score	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

Evaluation Task

- Lemmatization and morphosyntactic information tagging
- 90% for training; and 10% for test (unless less than 50 inflection tables)
- The evaluation data is **disjoint** from the training data
- The first-ranked analyses
- Recall
  - lemma
  - lemma + POS
  - lemma + MST

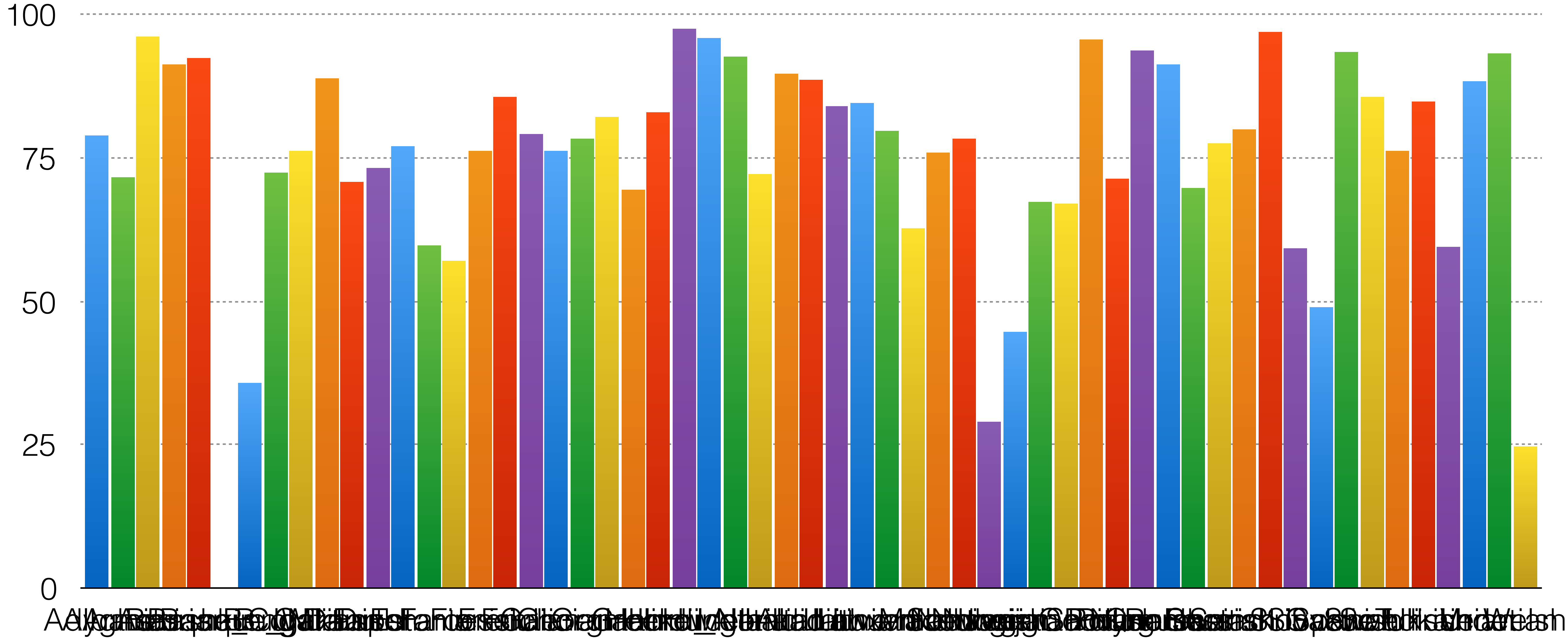


rank	score	paradigm	variables	lemma	mst
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;PST]**
1	-11.44	p4_unmarry	(1=verif)	**verify**	**[V;V.PTCP;PST]**
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;PST]
2	-18.36	p1_dribble	(1=verifi)	verifie	[V;V.PTCP;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;PST]
3	-30.49	p20_preempt	(1=verifi)	verifi	[V;V.PTCP;PST]

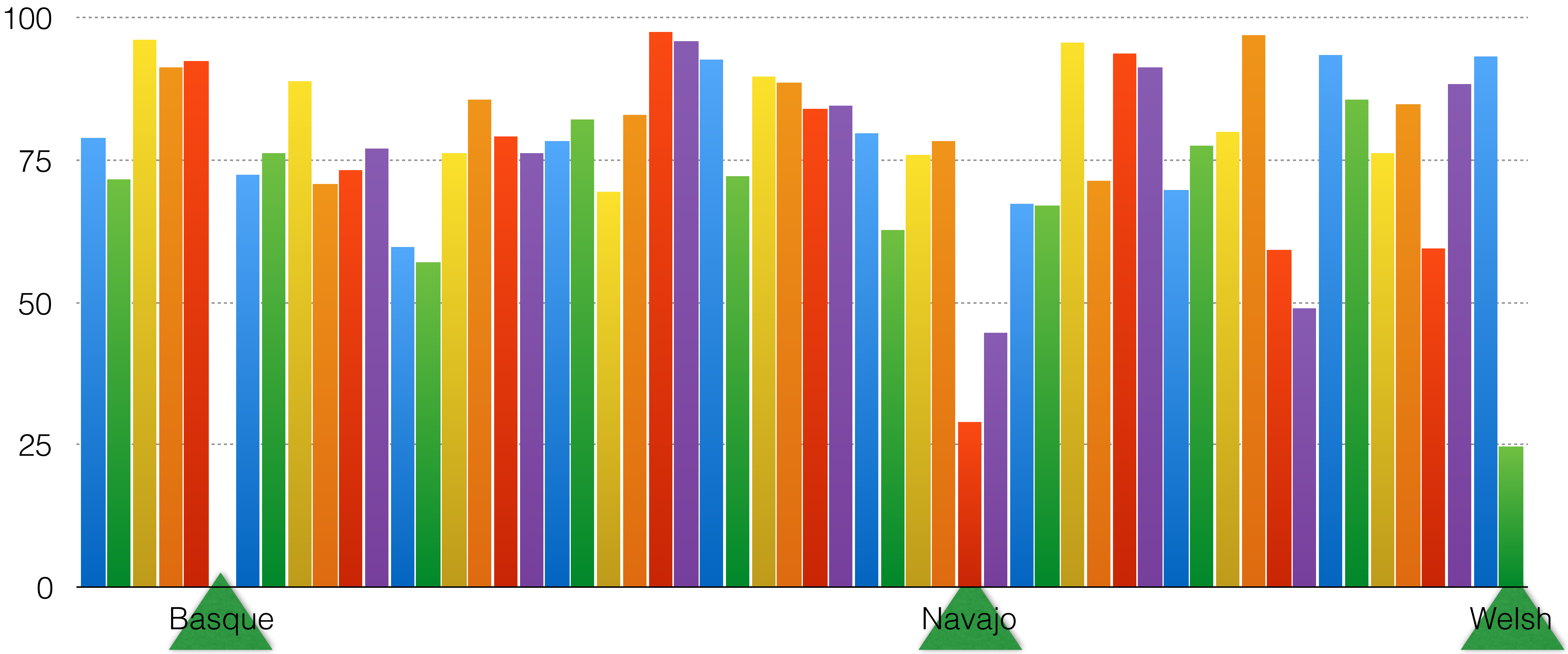
Result

- Paradigms are extracted successfully for all languages.
- **Lemmatization recall:**
  - Low end: 0% (Basque)
  - High end: 97.5% (Hindi)
- **Lemma-POS recall:**
  - Low end: 0% (Basque)
  - High end: 97.0% (Hindi)
- **Lemma-tag recall:**
  - Low end: 0% (Basque)
  - High end: 96.9% (Hindi)

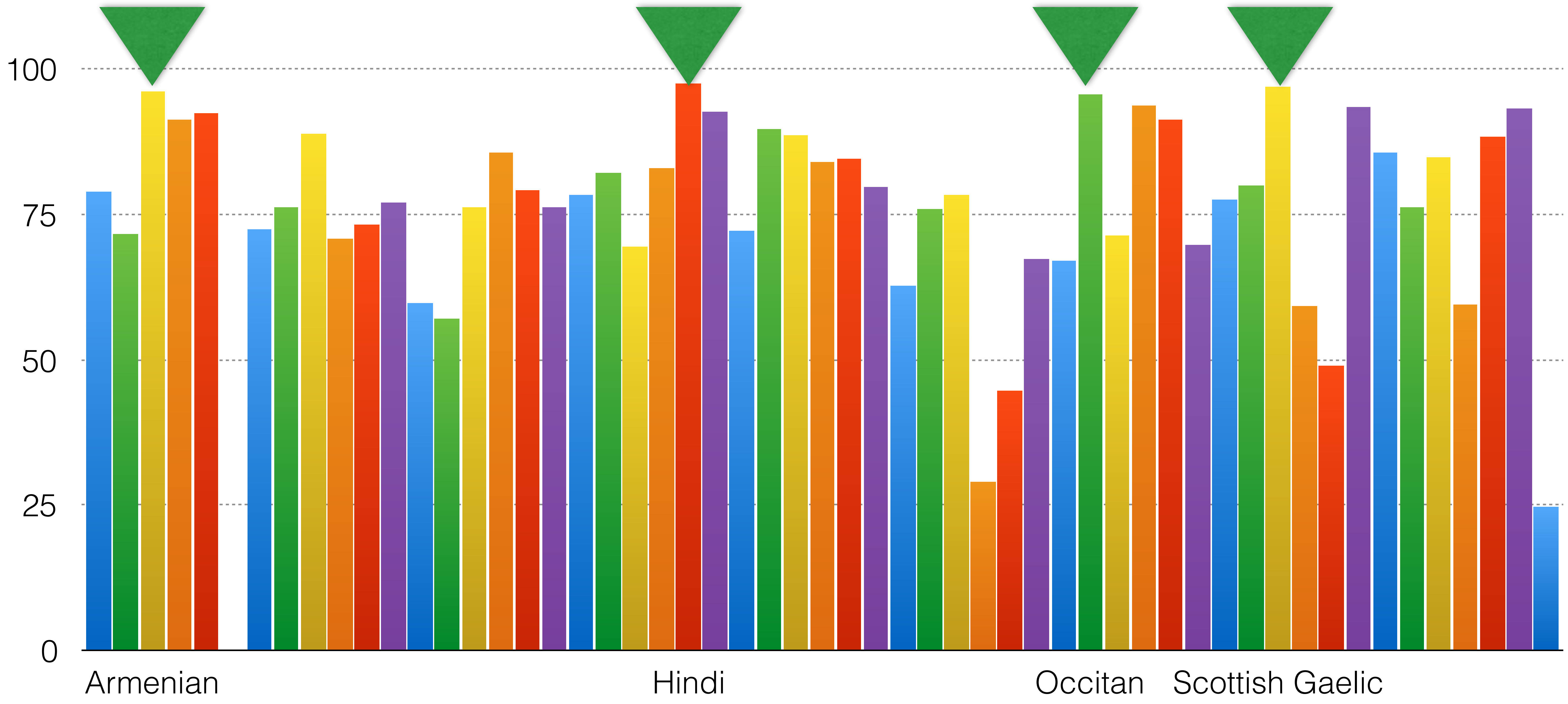
Result Overview: Lemma Recall



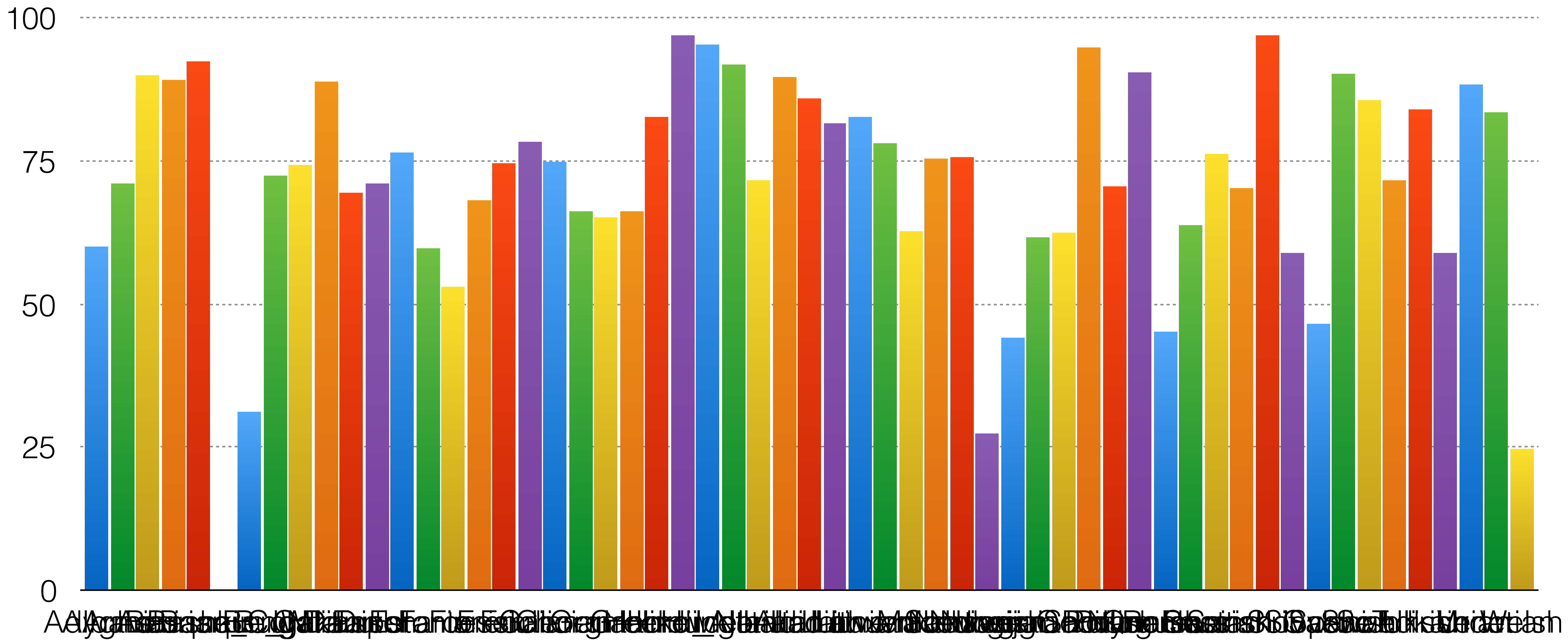
Result Overview: Lemma Recall < 30%



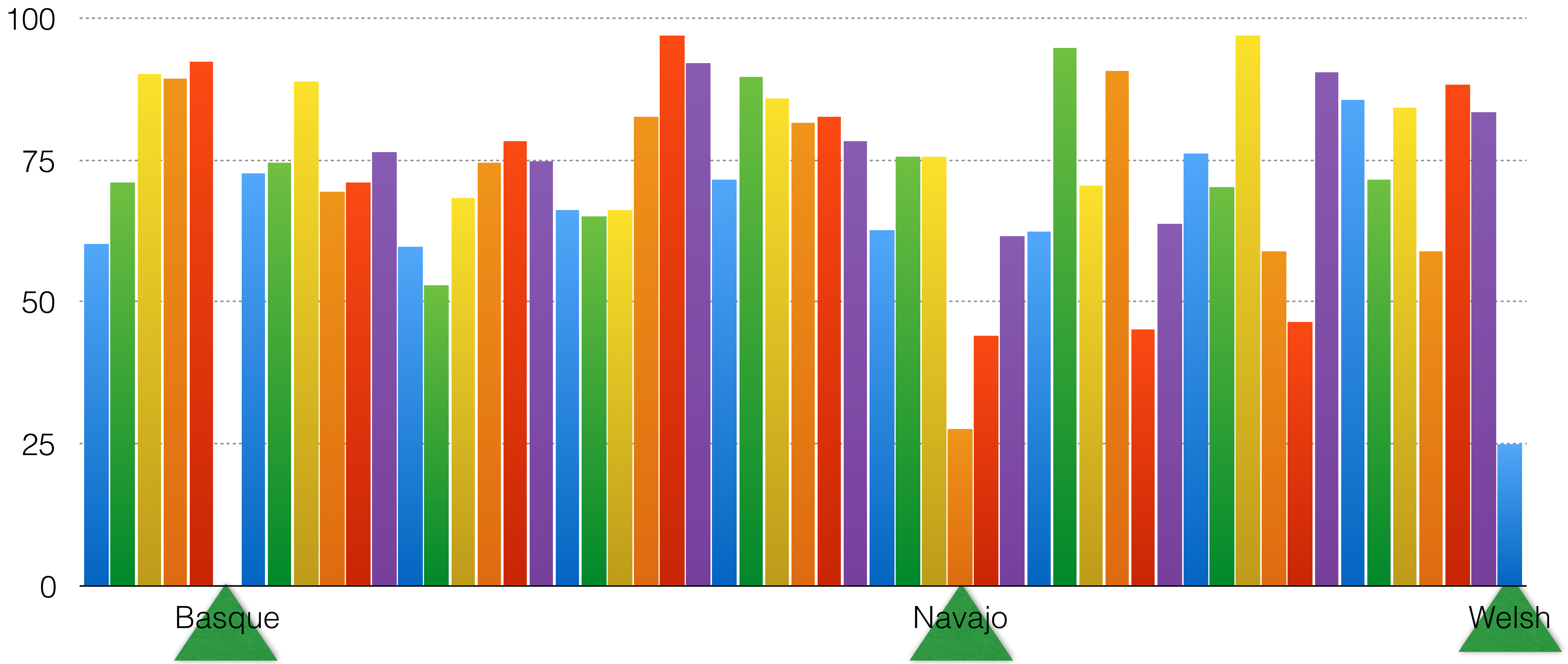
Result Overview: Lemma Recall > 95%



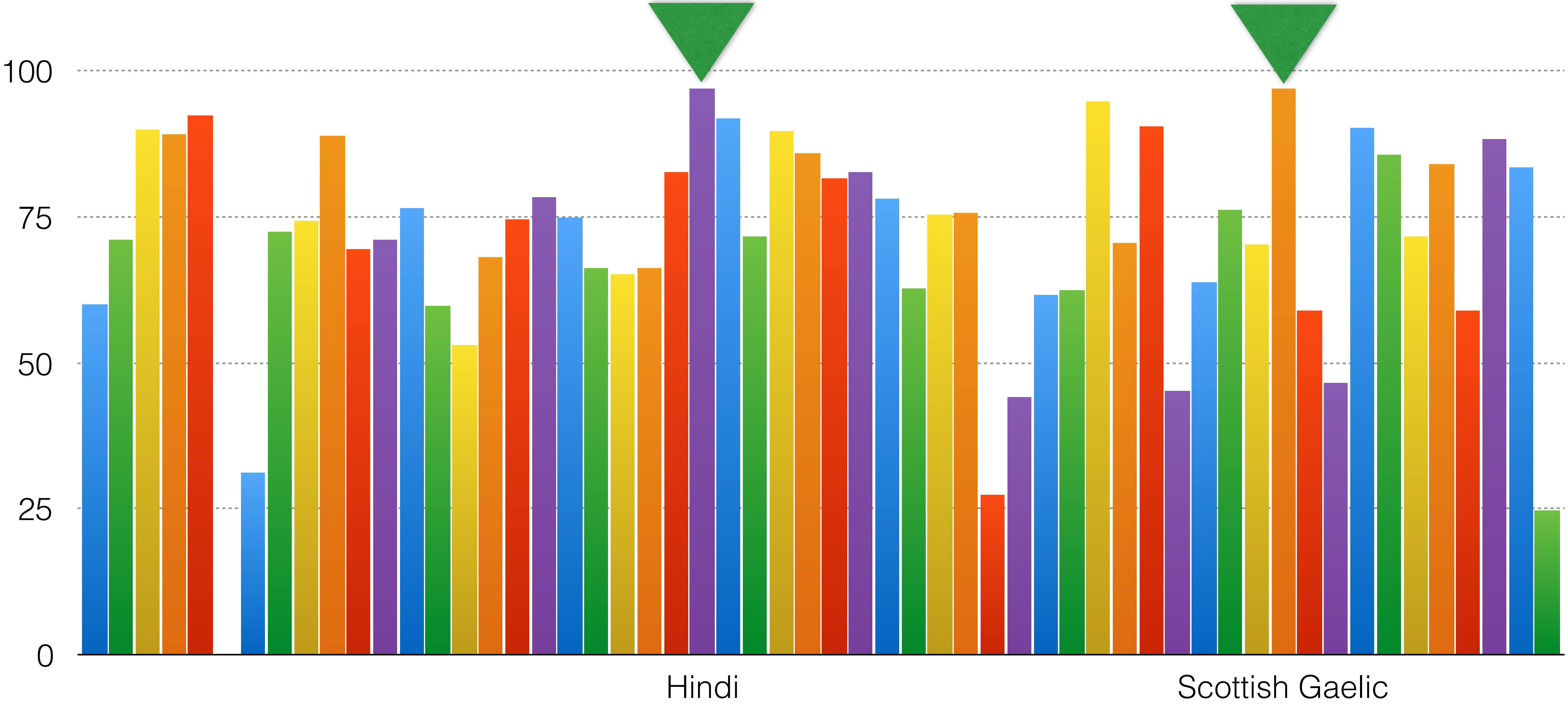
Result Overview: Lemma+POS Recall



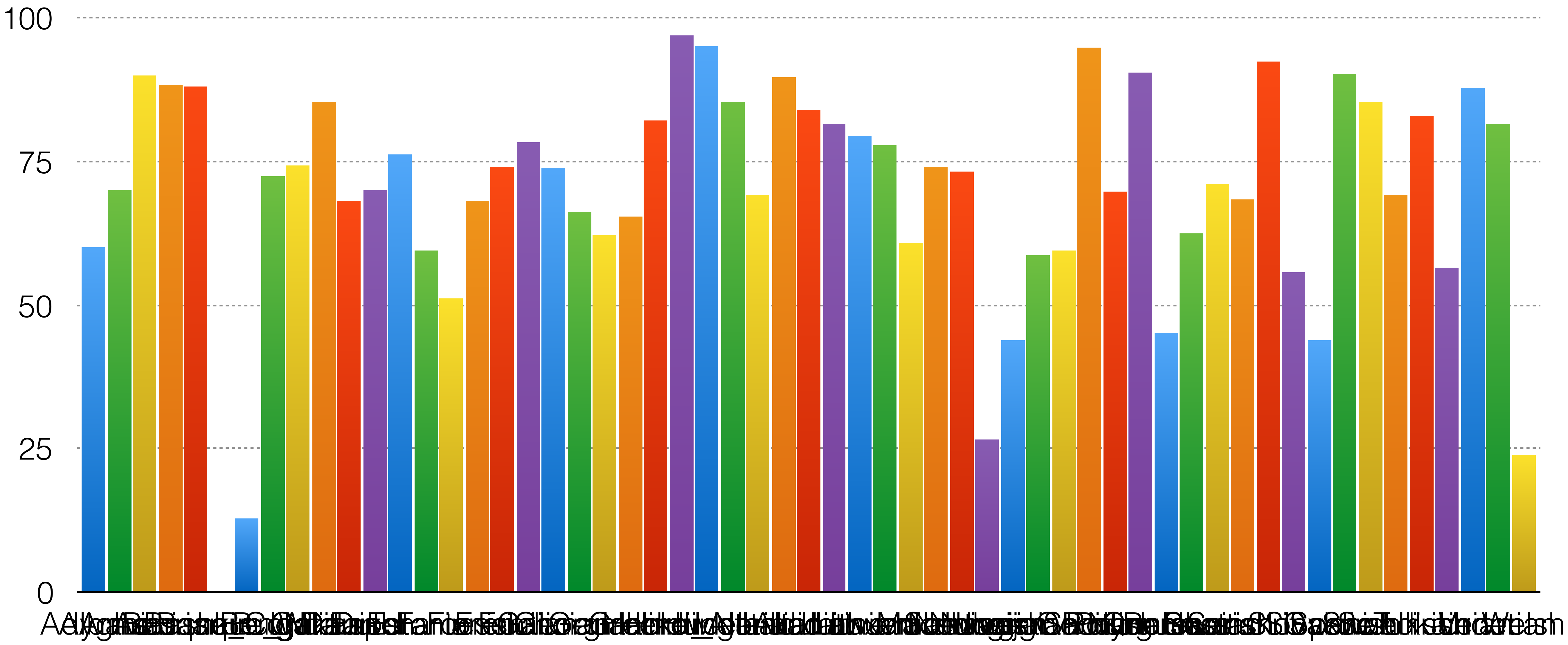
Result Overview: Lemma+POS Recall < 30%



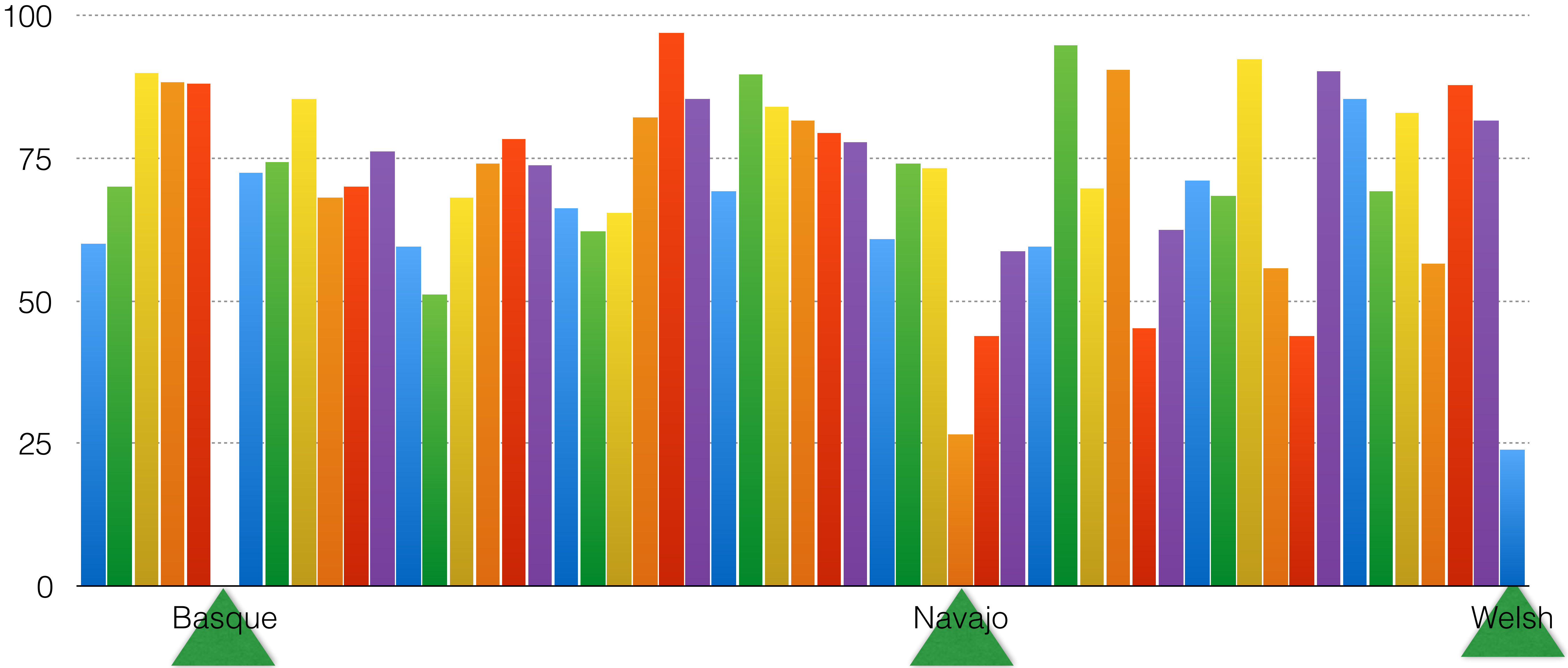
Result Overview: Lemma+POS Recall > 95%



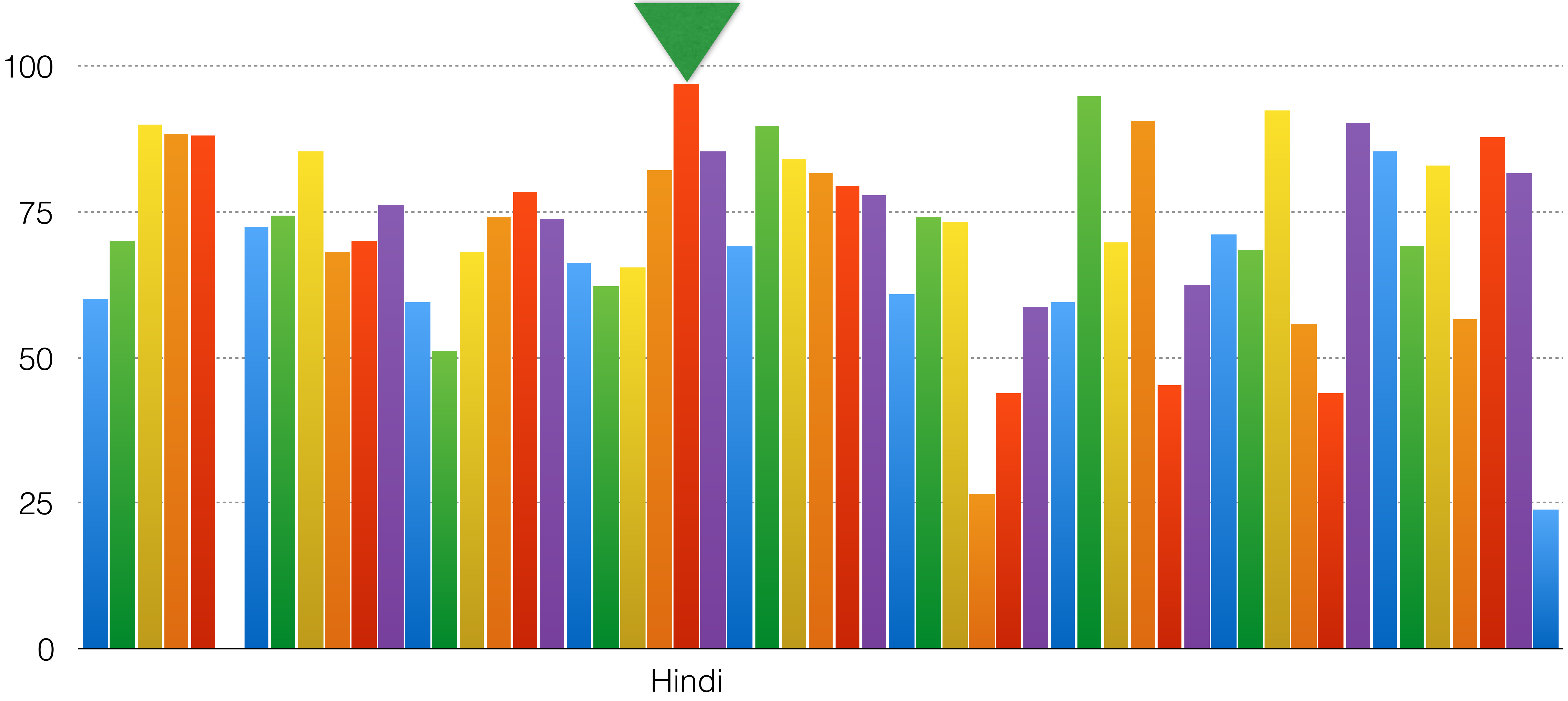
Result Overview: Lemma+Tags Recall



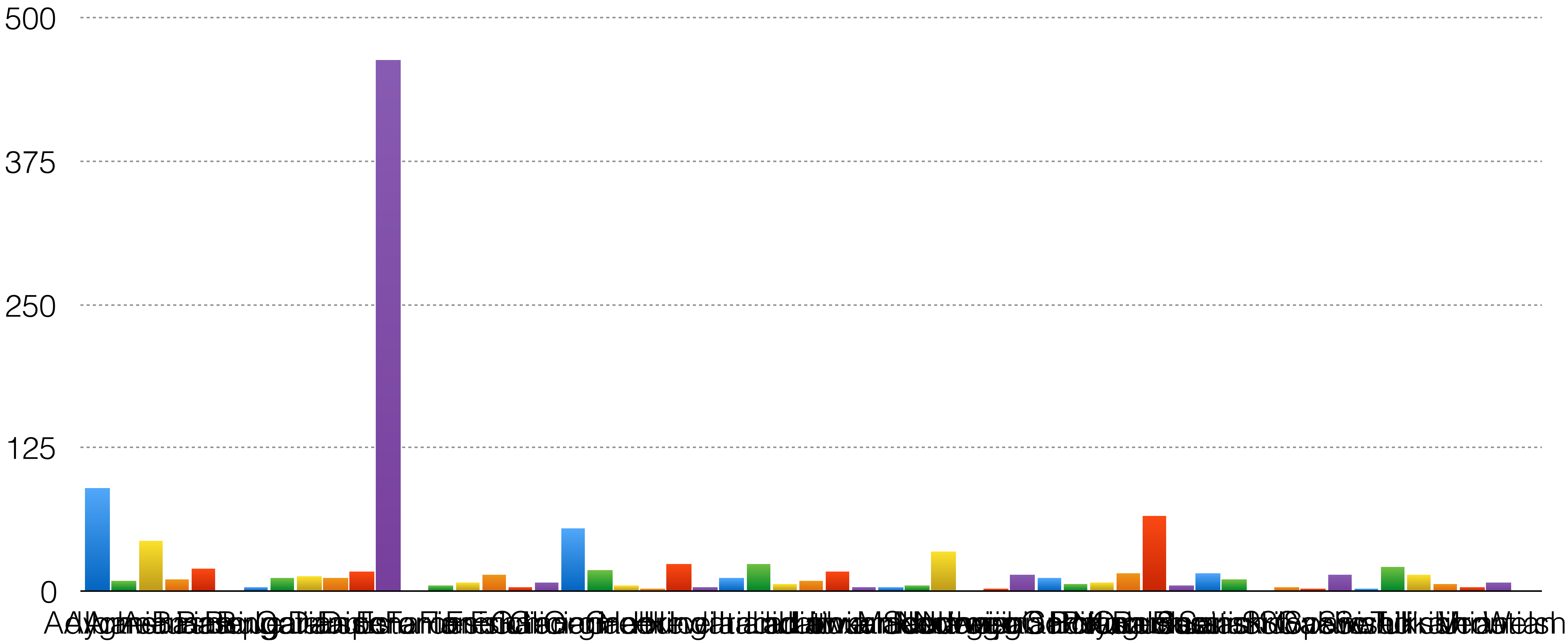
Result Overview: Lemma+Tags Recall < 30%



Result Overview: Lemma+Tags Recall > 95%

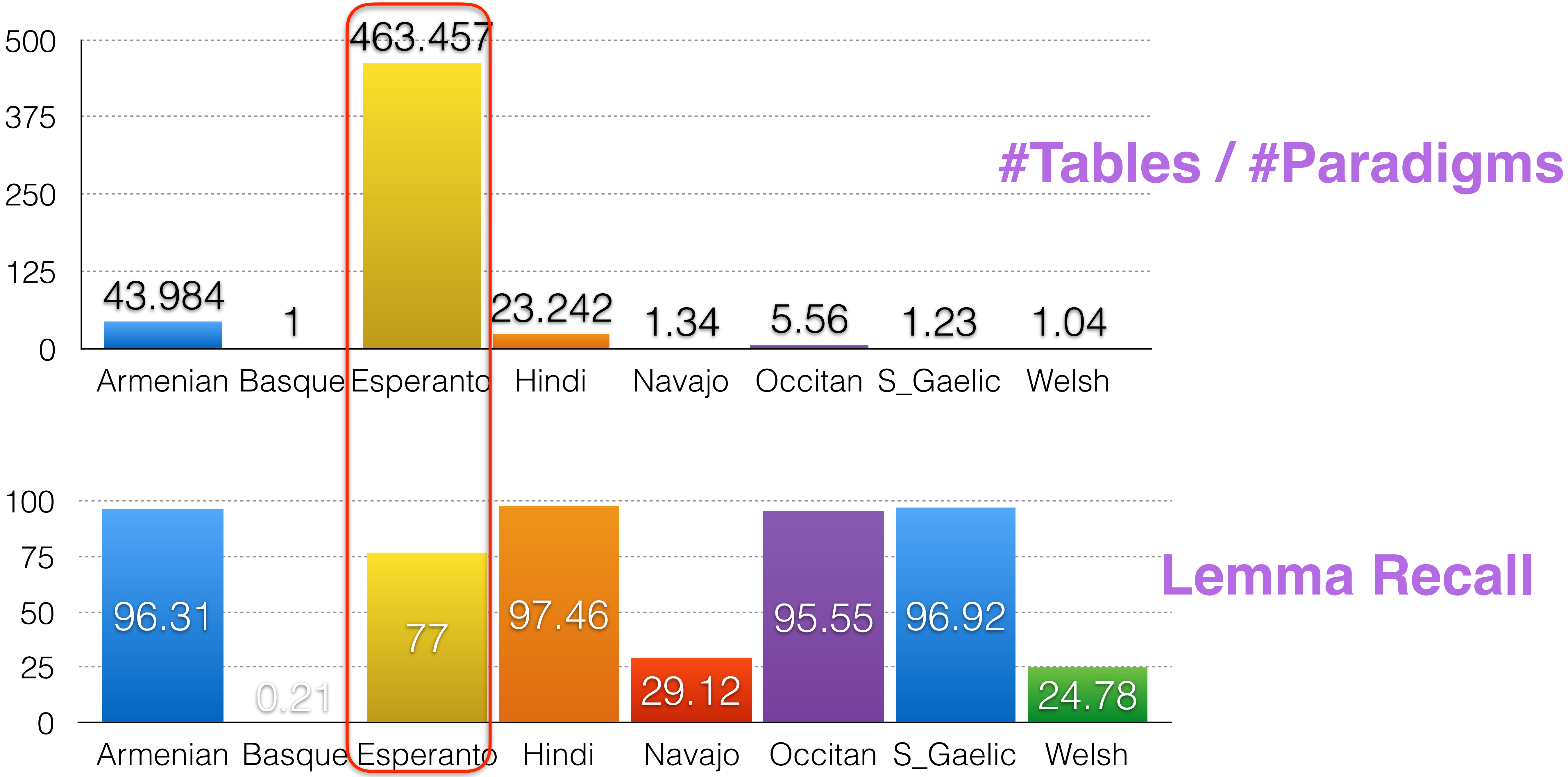


Overview: #Tables/#Paradigms



Result

- The results seem to correlate strongly with the amount and representativeness of available data.

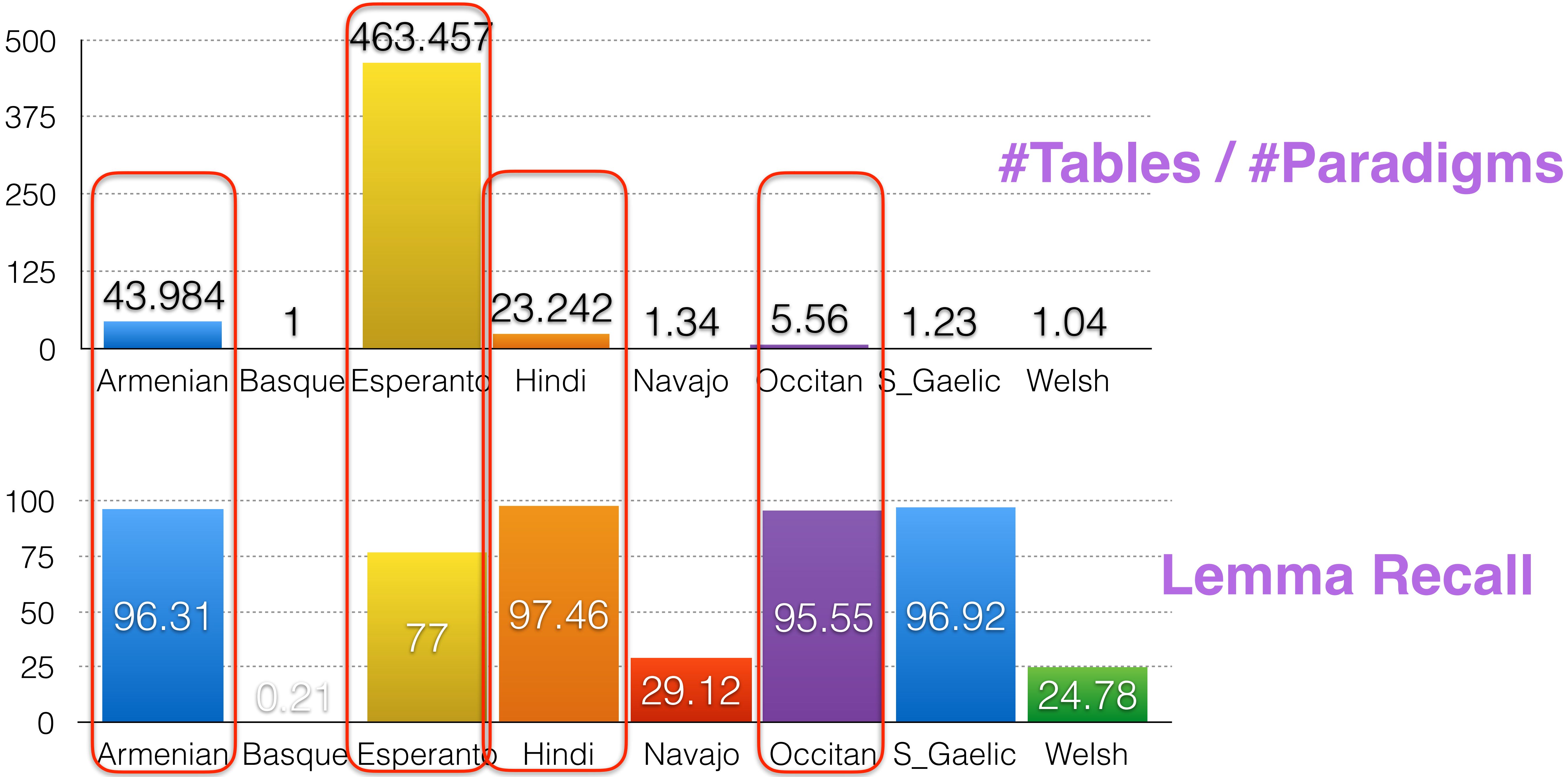


#Tables / #Paradigms

Lemma Recall

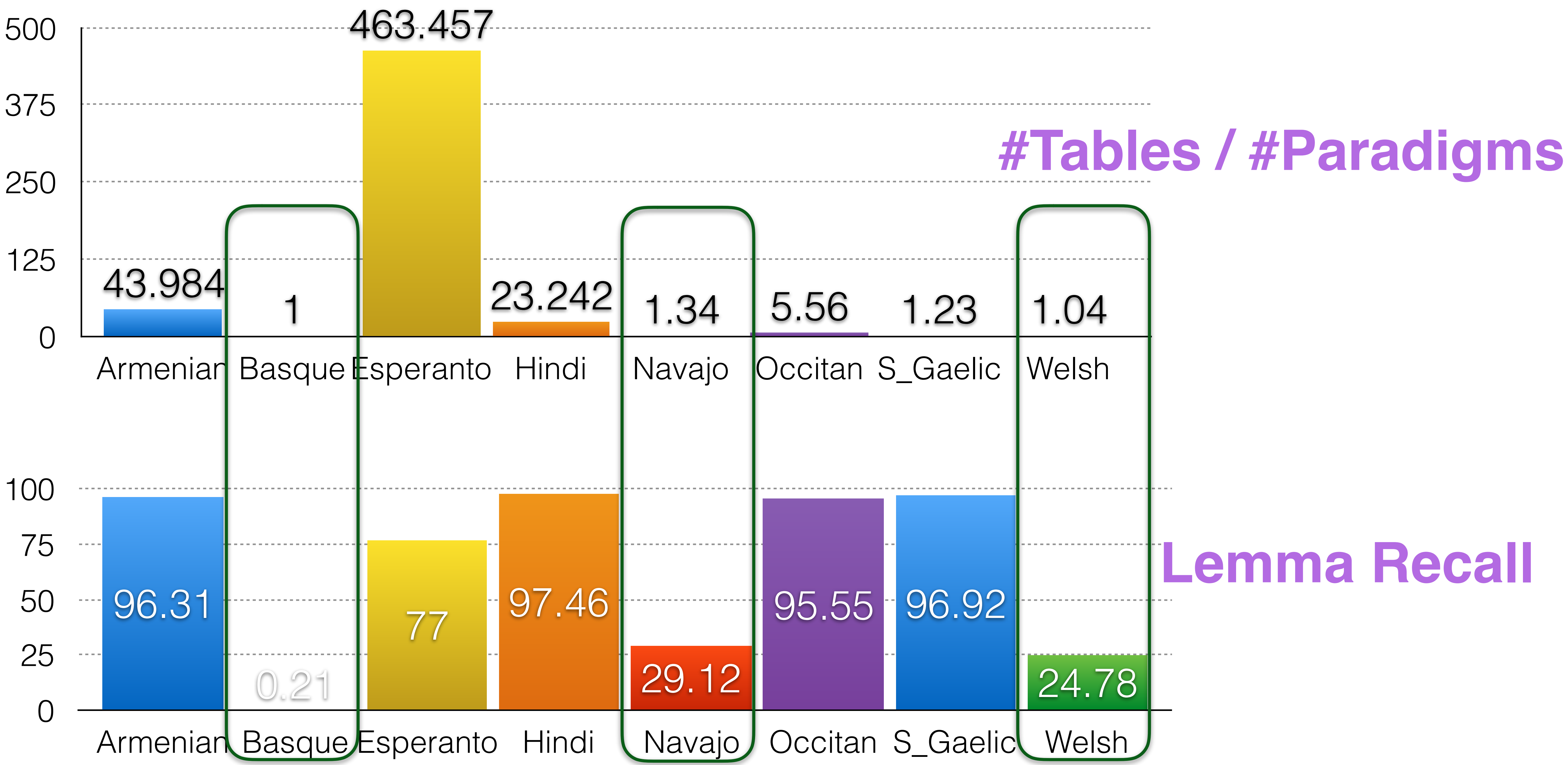
Result

- The results seem to correlate strongly with the amount and representativeness of available data.



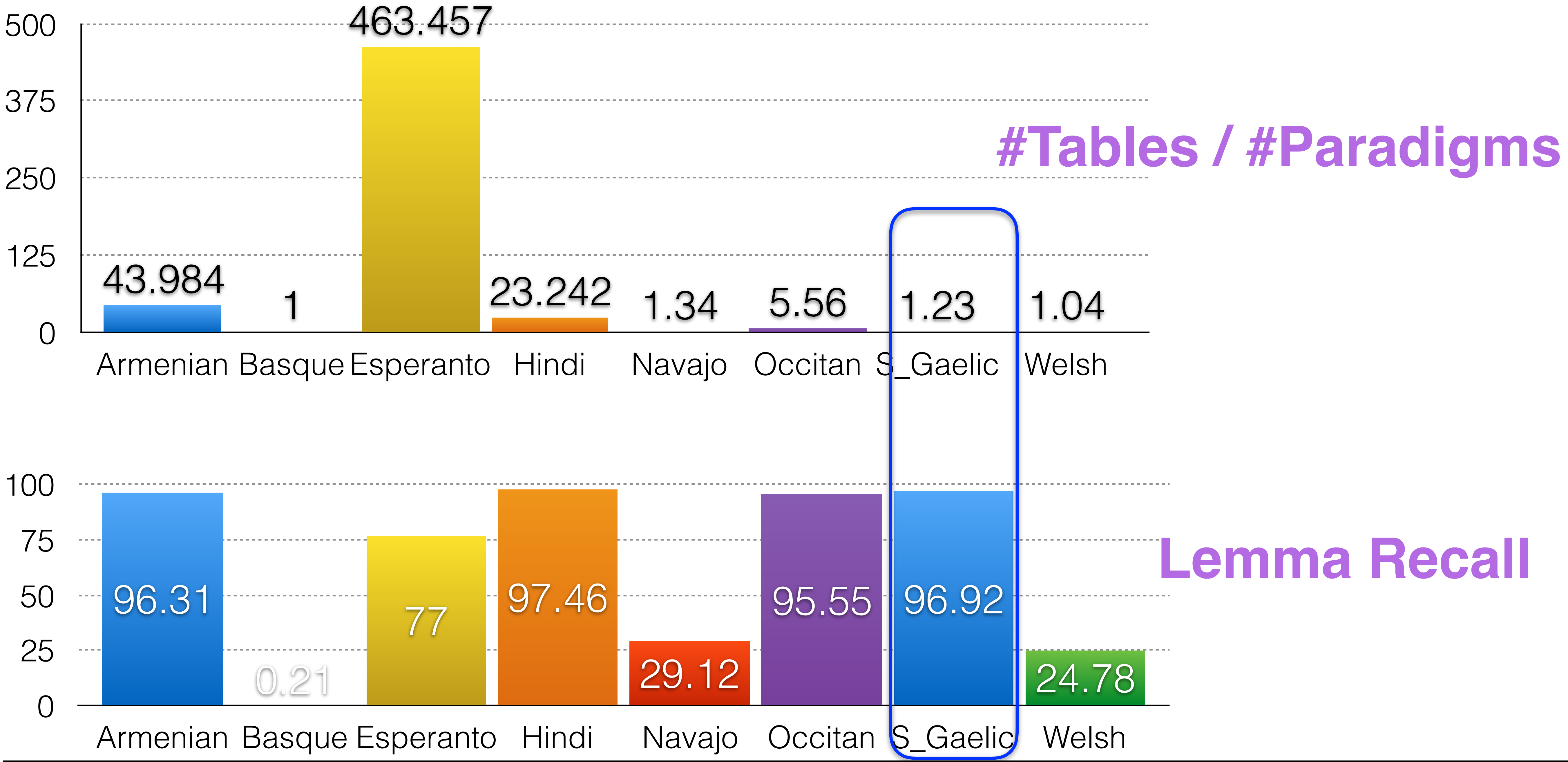
Result

- The results seem to correlate strongly with the amount and representativeness of available data.



Result

- The results seem to correlate strongly with the amount and representativeness of available data.



#Tables / #Paradigms

Lemma Recall

Wrap-up

- Simple method to construct weighted FST from labeled data
- Robust performance for inflectional morphology
- Large representative data is critical for the performance.

Thank You