

Olá, Bonjour, Salve!

XFORMAL: A Benchmark for Multilingual Formality Style Transfer

Eleftheria Briakou, Di Lu, Ke Zhang, Joel Tetreault



Formality Style



“style is an intuitive notion involving the manner in which something is said”

McDonald and Pustejovsky. 1985



“a dimension similar to **formality** appears as the most **important** and **universal** feature distinguishing styles, registers or genres **in different languages**”

Heylighen and Dewaele, 1999

Formality Style



“style is an intuitive notion involving the manner in which something is said”

McDonald and Pustejovsky. 1985



“a dimension similar to formality appears as the most important and universal feature distinguishing styles, registers or genres in different languages”

Heylighen and Dewaele, 1999

Formality Style



“style is an intuitive notion involving the manner in which something is said”

McDonald and Pustejovsky. 1985



“a dimension similar to **formality appears as the most **important** and **universal** feature distinguishing styles, registers or genres **in different languages**”**

Heylighen and Dewaele, 1999

Formality Style Transfer (FoST): Task Definition

Generate a well-formed sentence that matches a desired formality attribute while preserving the meaning of the input

Formality Style Transfer (FoST): Task Definition

Generate a well-formed sentence that matches a desired formality attribute while preserving the meaning of the input

INFORMAL

Gotta see both sides of the story

Formality Style Transfer (FoST): Task Definition

Generate a well-formed sentence that matches a desired formality attribute while preserving the meaning of the input

INFORMAL

Gotta see both sides of the story

FORMAL

You have to consider both sides of the story

Formality Style Transfer: Current status

Hello!

Olá

Bonjour

Salve

Formality Style Transfer: Current status

GYAFC [Rao & Tetreault; 2018]

Hello!

Olá

Bonjour

Salve

First large scale dataset
of informal-formal parallel pairs in En

Formality Style Transfer: This work

GYAFC [Rao & Tetreault; 2018]

Hello!

Olá

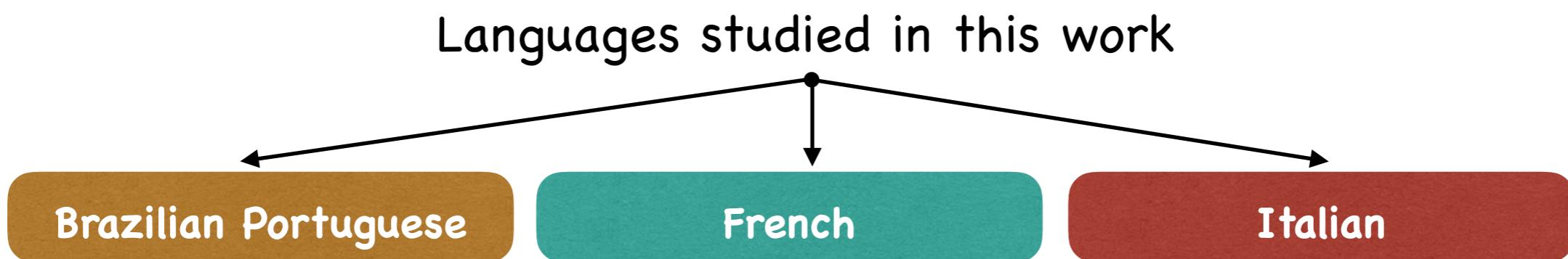
Bonjour

Salve

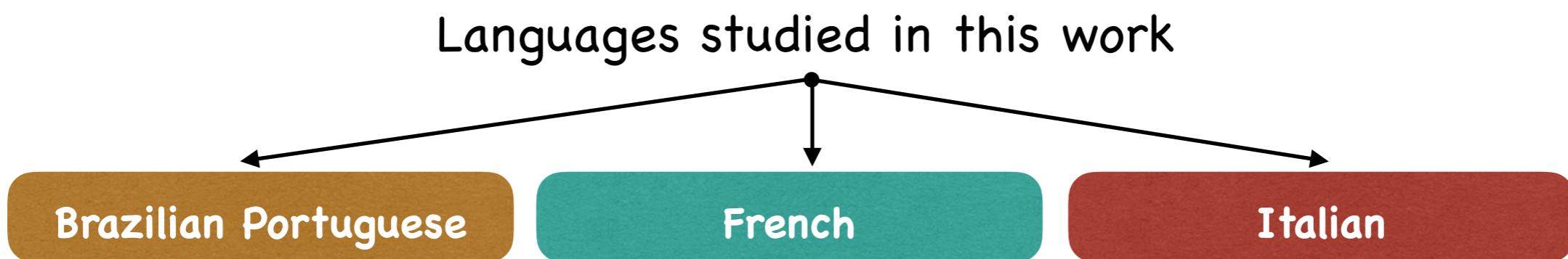
First large scale dataset
of informal-formal parallel pairs in En

How well can we perform formality style transfer in
different languages?

How well can we perform formality style transfer in different languages?

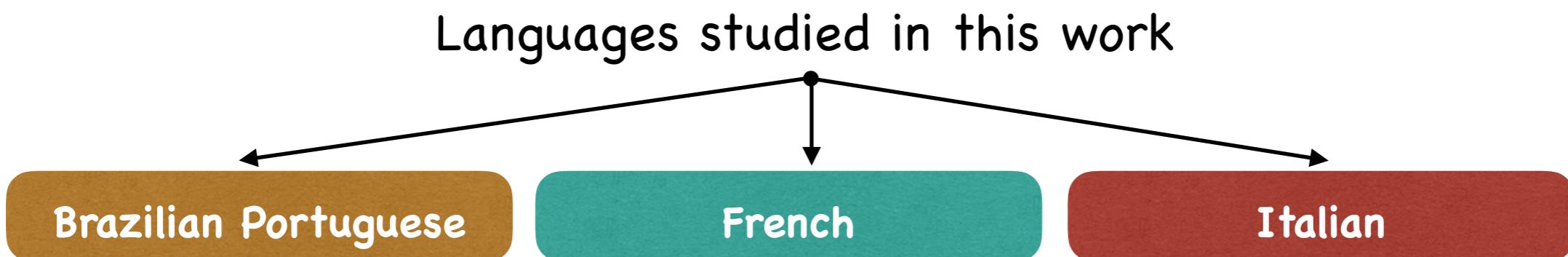


How well can we perform formality style transfer in different languages?



First multilingual FoST set ever!
Successful crowd-sourcing w/ Careful Quality Control
Head to head comparison of FoST systems across 3 languages

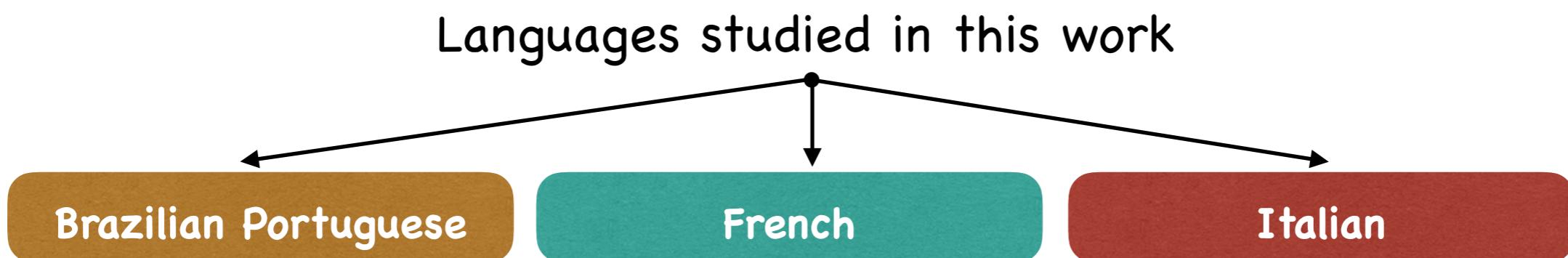
How well can we perform formality style transfer in different languages?



Main challenge: Availability of resources

- for evaluation
- for training

How well can we perform formality style transfer in different languages?



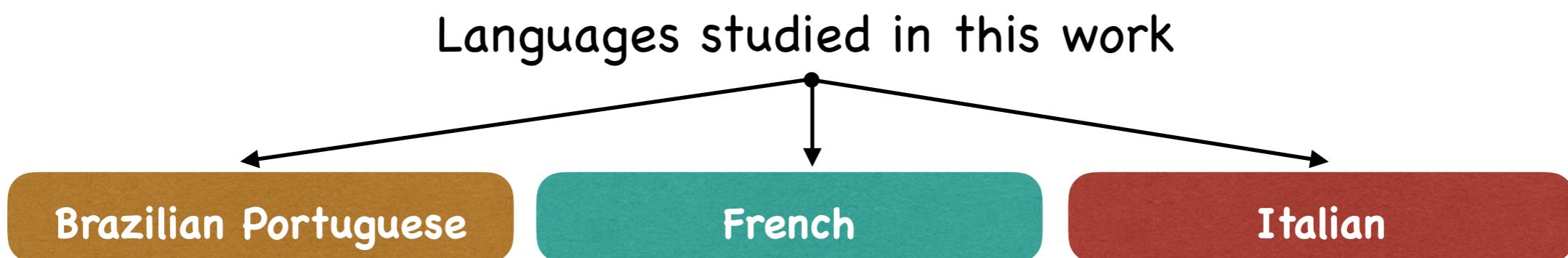
Main challenge: Availability of resources

- for evaluation
- for training

XFORMAL

Informal excerpts paired
with formal rewrites

How well can we perform formality style transfer in different languages?



Main challenge: Availability of resources

- for evaluation
- for training

XFORMAL

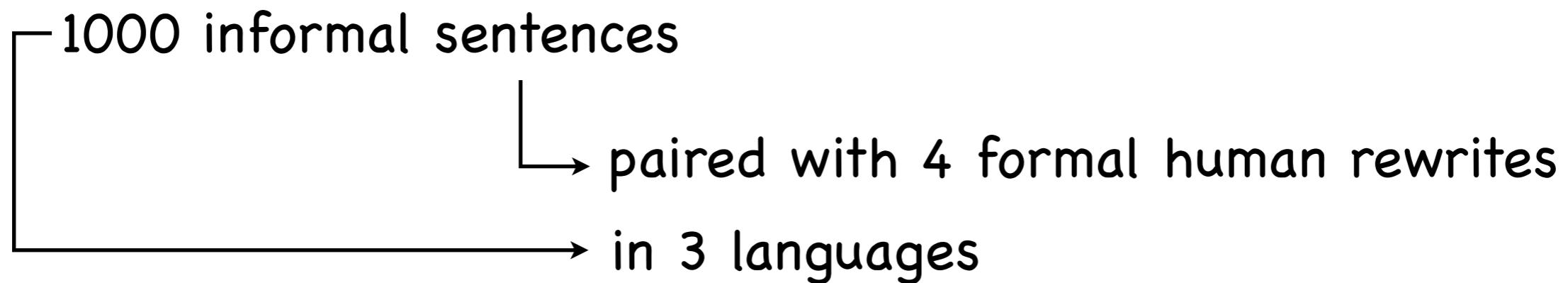
cross-lingual transfer
synthetic supervision
unsupervised

Informal excerpts paired
with formal rewrites

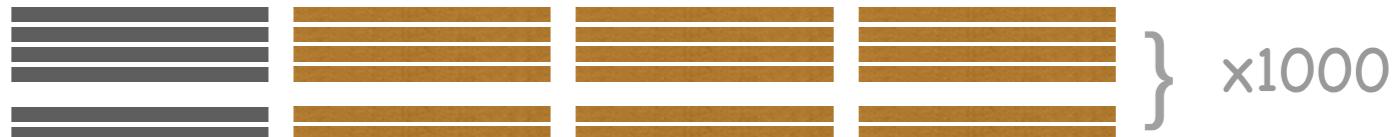
Annotation is hard to scale!

Introducing XFORMAL: Data description

Evaluation dataset for multilingual formality style transfer



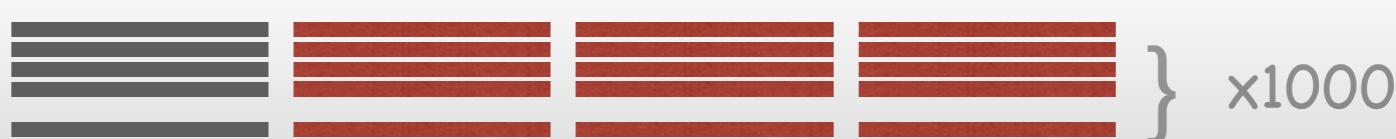
Brazilian Portuguese



French



Italian



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control

Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control

L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)

Yahoo! Answers is a web site where people post questions and answers, all of which are public to any web user willing to browse or download them. The data we have collected is the Yahoo! Answers corpus as of 10/25/2007. It includes all the questions and their corresponding answers. The corpus distributed here contains 4,483,032 questions and their answers. In addition to question and answer text, the corpus contains a small amount of metadata, i.e., which answer was selected as the best answer, and the category and sub-category that was assigned to this question. No personal information is included in the corpus. The question URIs and all user ids were anonymized so that no identifying information is revealed. This dataset may be used by researchers to learn and validate answer extraction models. An example of such work was published by Surdeanu et al. (2008). There are 2 files in this dataset. Part 1 is 1.7 Gbyte and part 2 is 1.9 Gybe.

Introducing XFORMAL: Data Collection

Curation rationale

Procedures

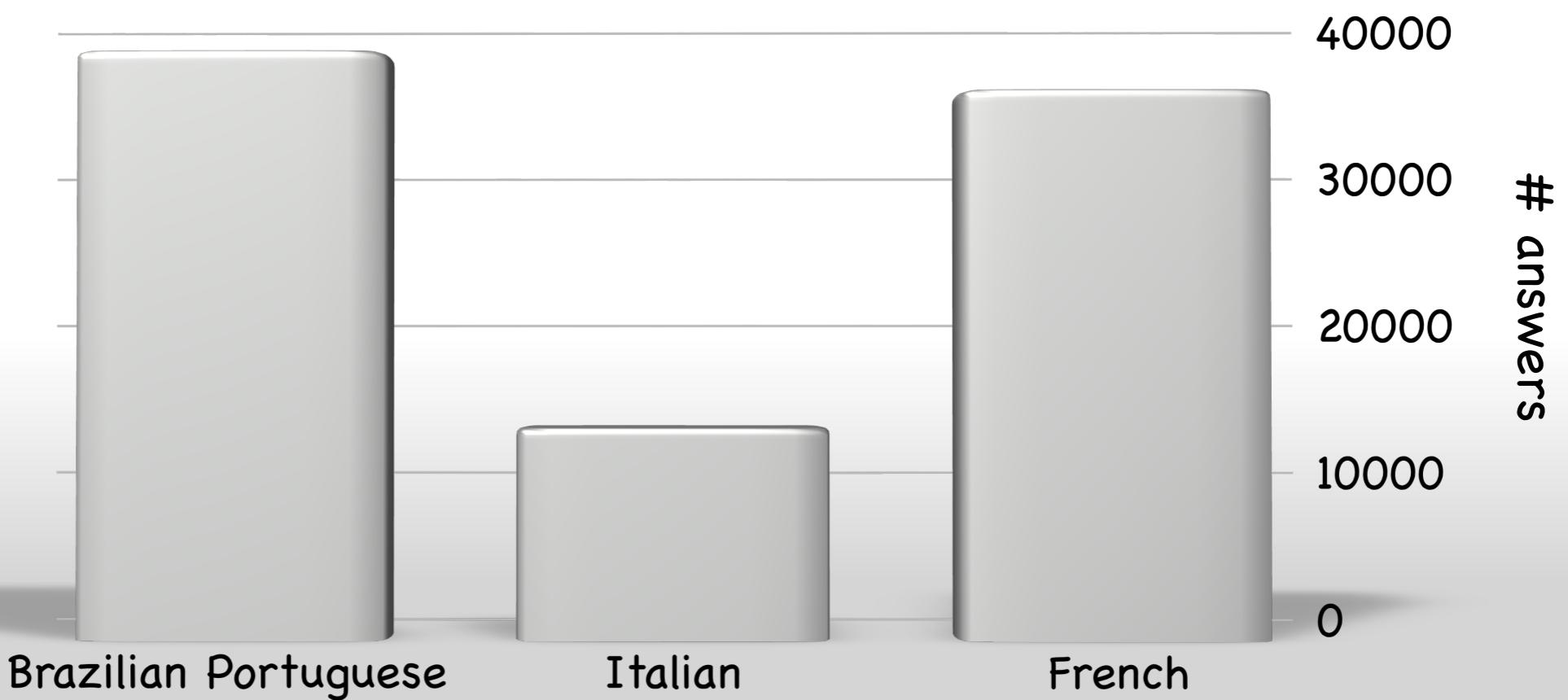
Quality Control

Step 1: Family & Relationship domain

Step 2: Pre-processing

Step 3: Detect informal answers

Step 4: Randomly sample 1000 answers



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

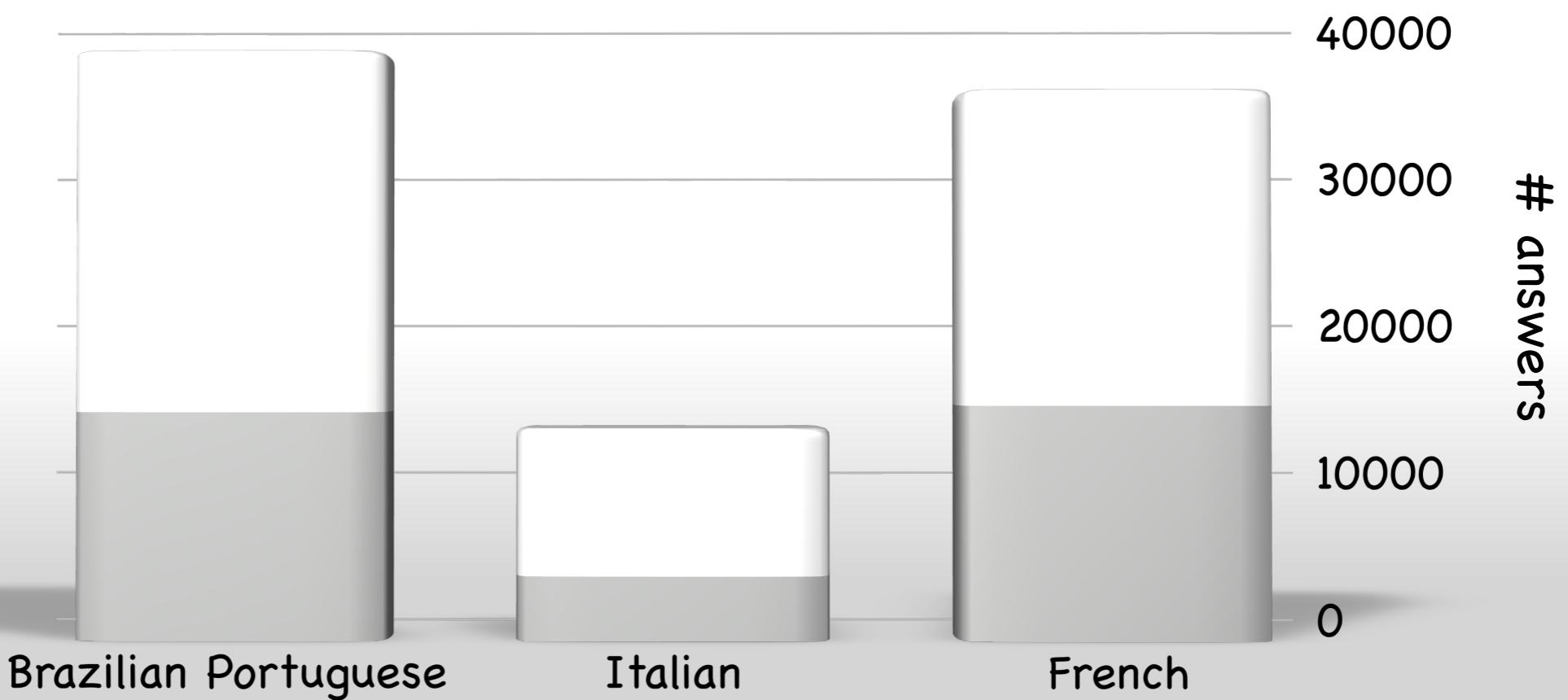
Quality Control

Step 1: Family & Relationship domain

Step 2: Pre-processing

Step 3: Detect informal answers

Step 4: Randomly sample 1000 answers



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

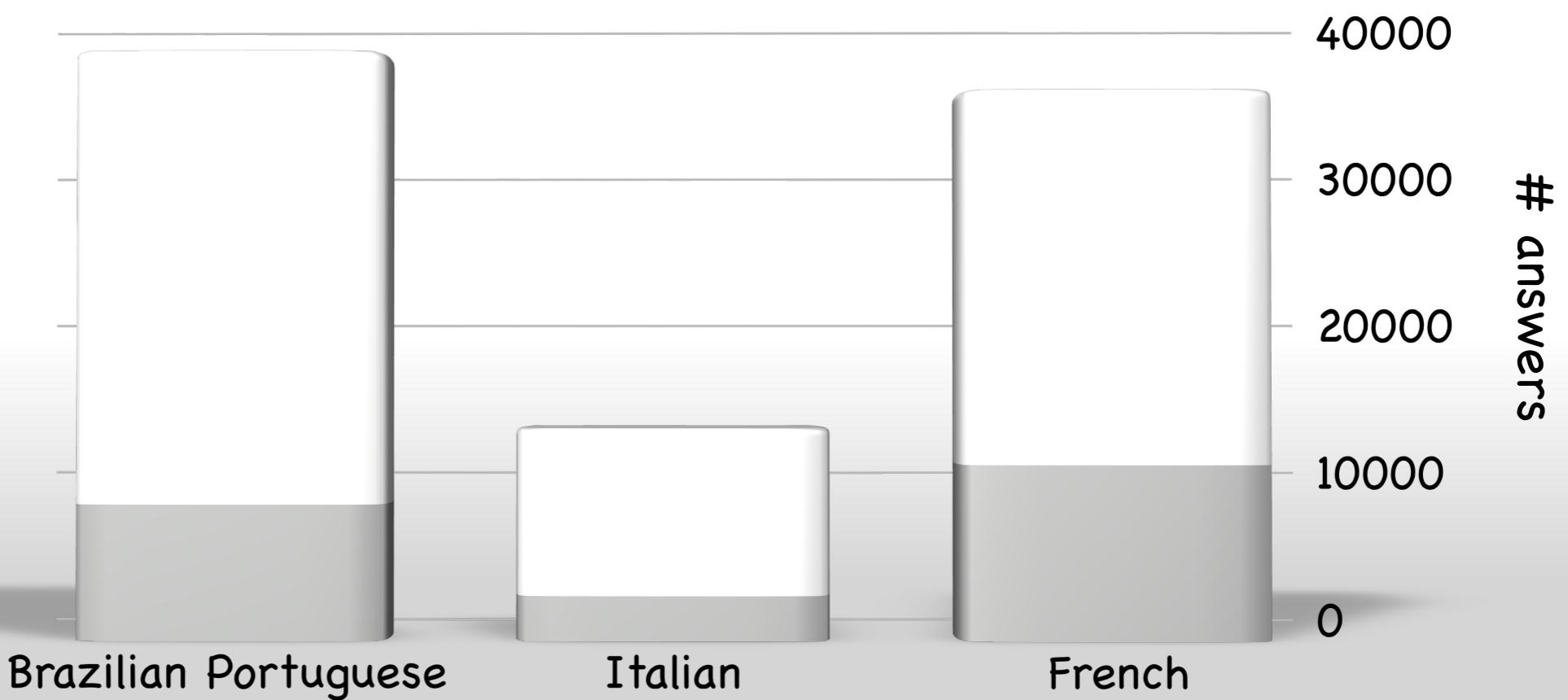
Quality Control

Step 1: Family & Relationship domain

Step 2: Pre-processing

Step 3: Detect informal answers

Step 4: Randomly sample 1000 answers



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

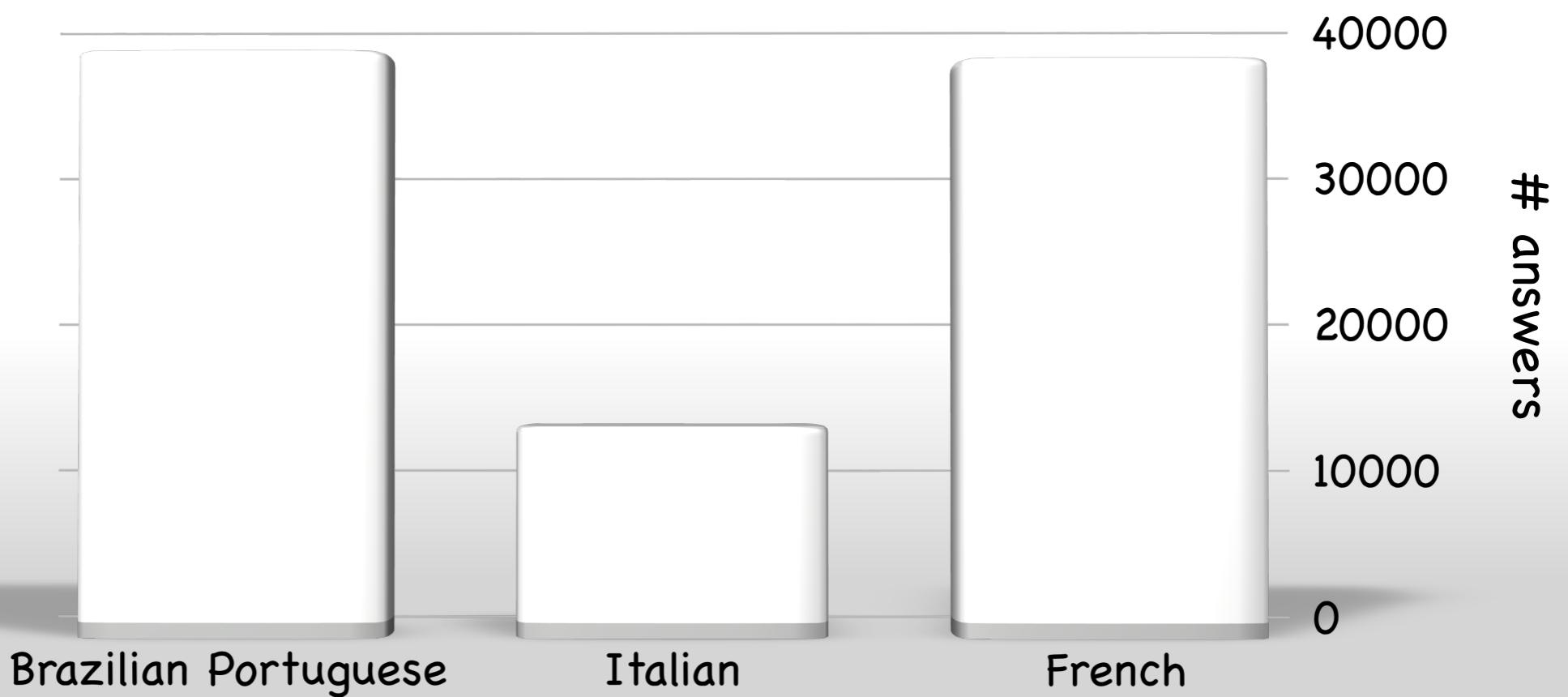
Quality Control

Step 1: Family & Relationship domain

Step 2: Pre-processing

Step 3: Detect informal answers

Step 4: Randomly sample 1000 answers



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control



Given an informal excerpt generate its

- formal rewrite
- in the same language
- without changing its meaning

Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control



Amazon Mechanical Turk workers (Turkers)

Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control



Amazon Mechanical Turk workers (Turkers)

Challenge!

Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control



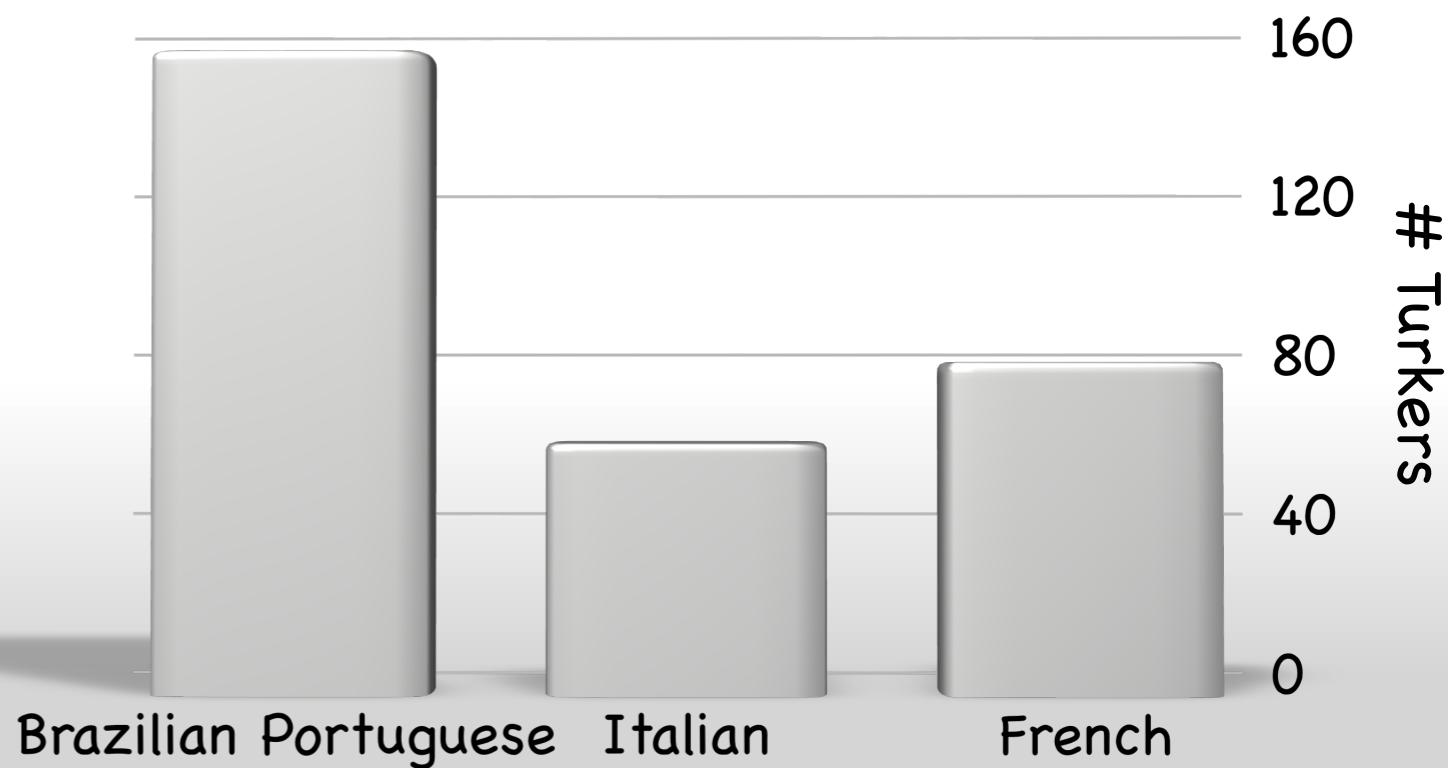
Amazon Mechanical Turk workers (Turkers)

Challenge!

QC1: Location Restrictions

QC2: Qualification test

QC3: Filtering by pilot study



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

Quality Control



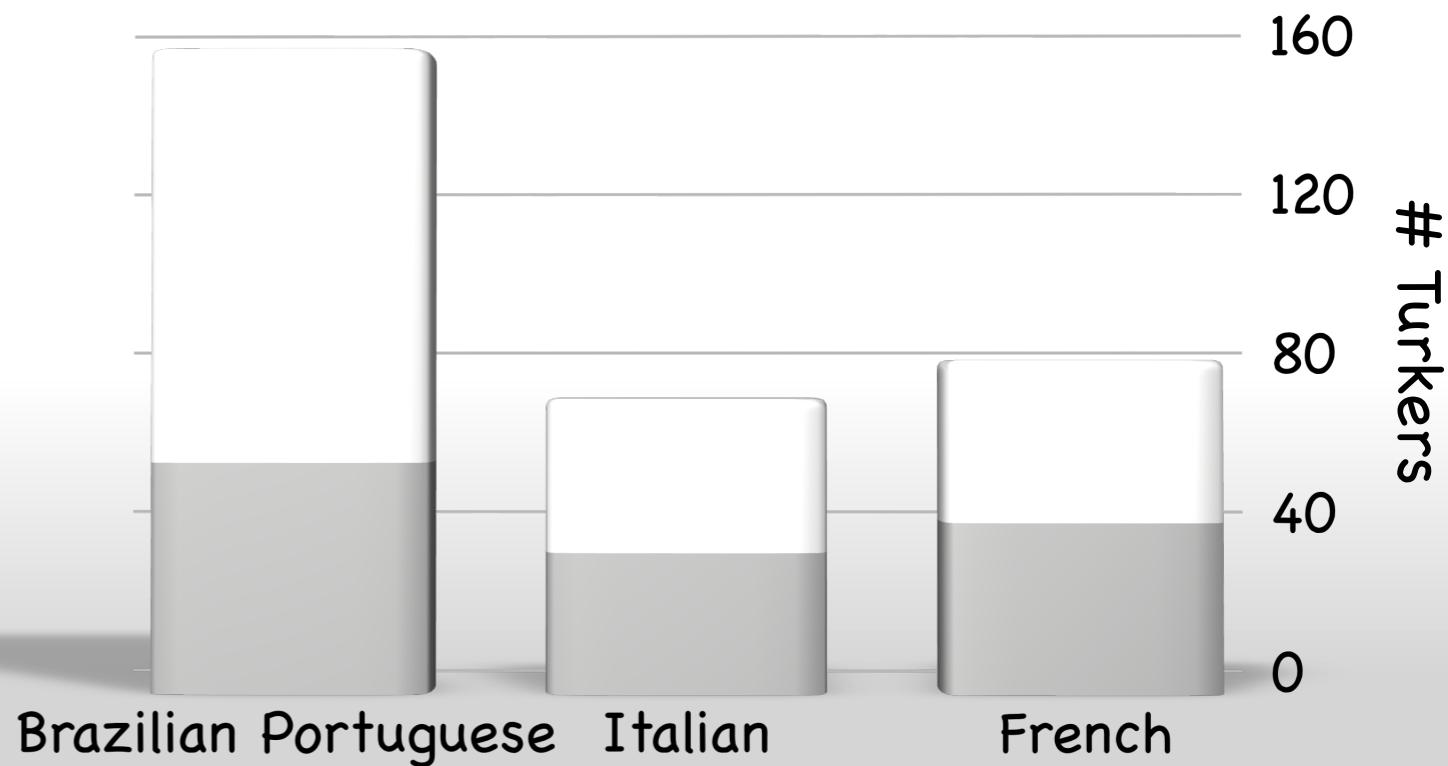
Amazon Mechanical Turk workers (Turkers)

Challenge!

QC1: Location Restrictions

QC2: Qualification test

QC3: Filtering by pilot study



Introducing XFORMAL: Data Collection

Curation rationale

Procedures

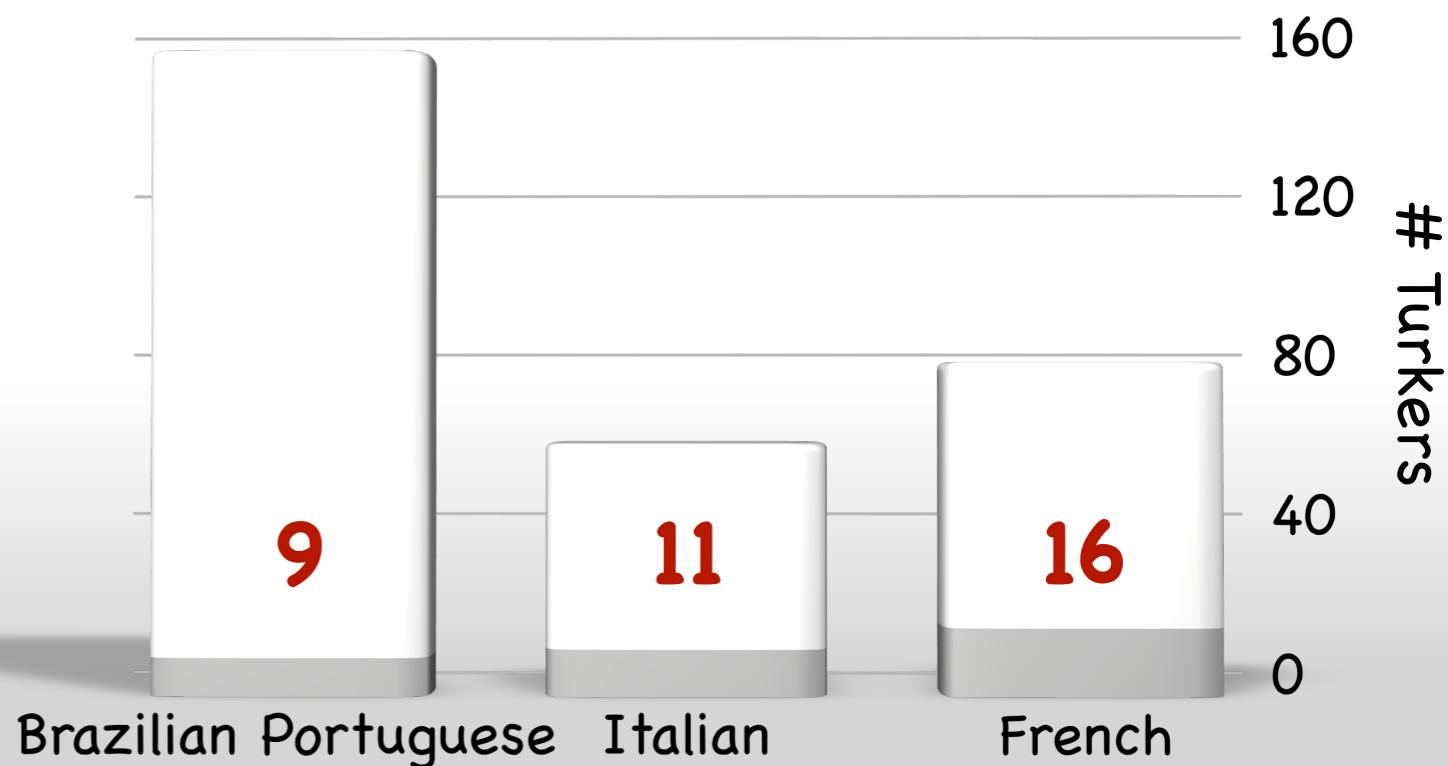
Quality Control



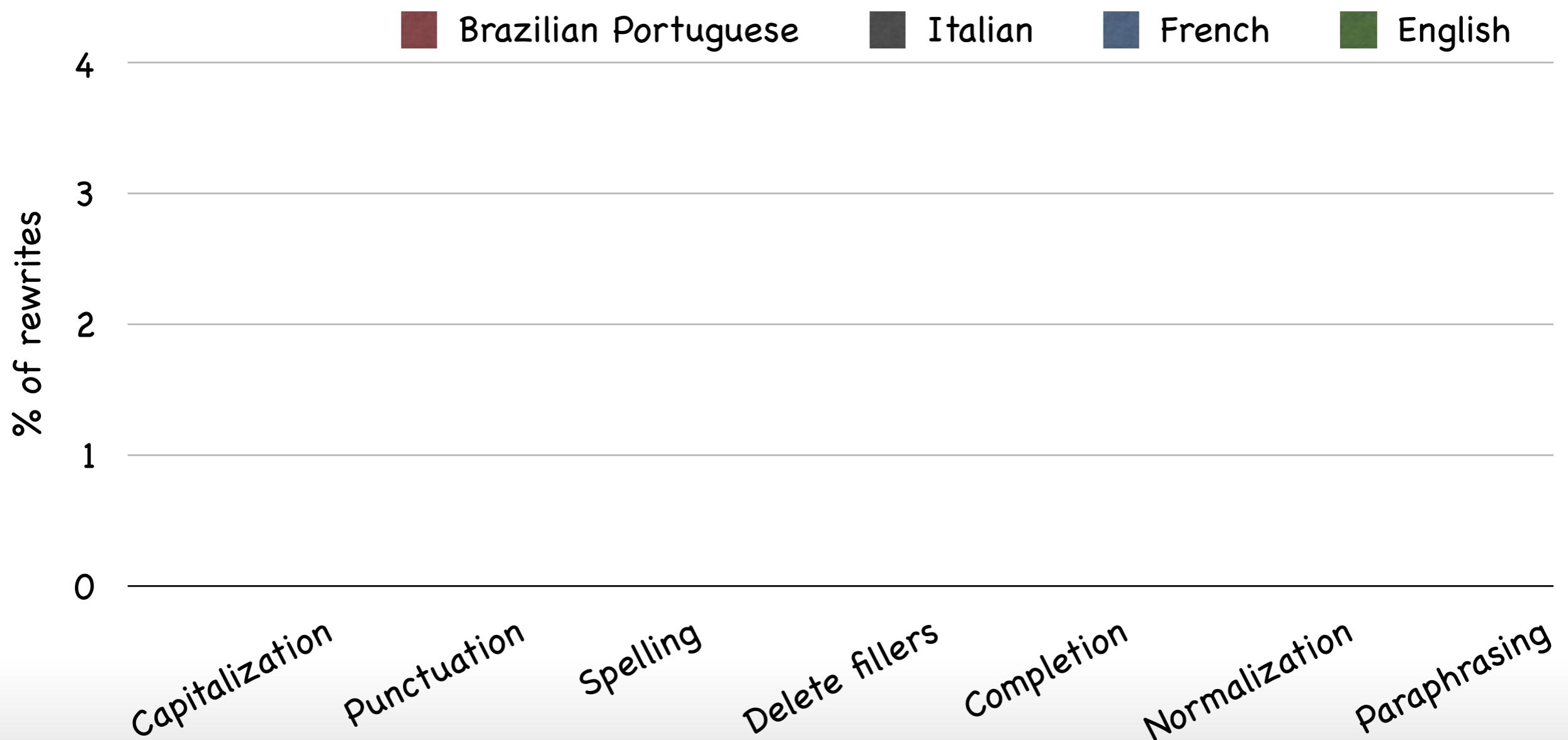
Amazon Mechanical Turk workers (Turkers)

Challenge!

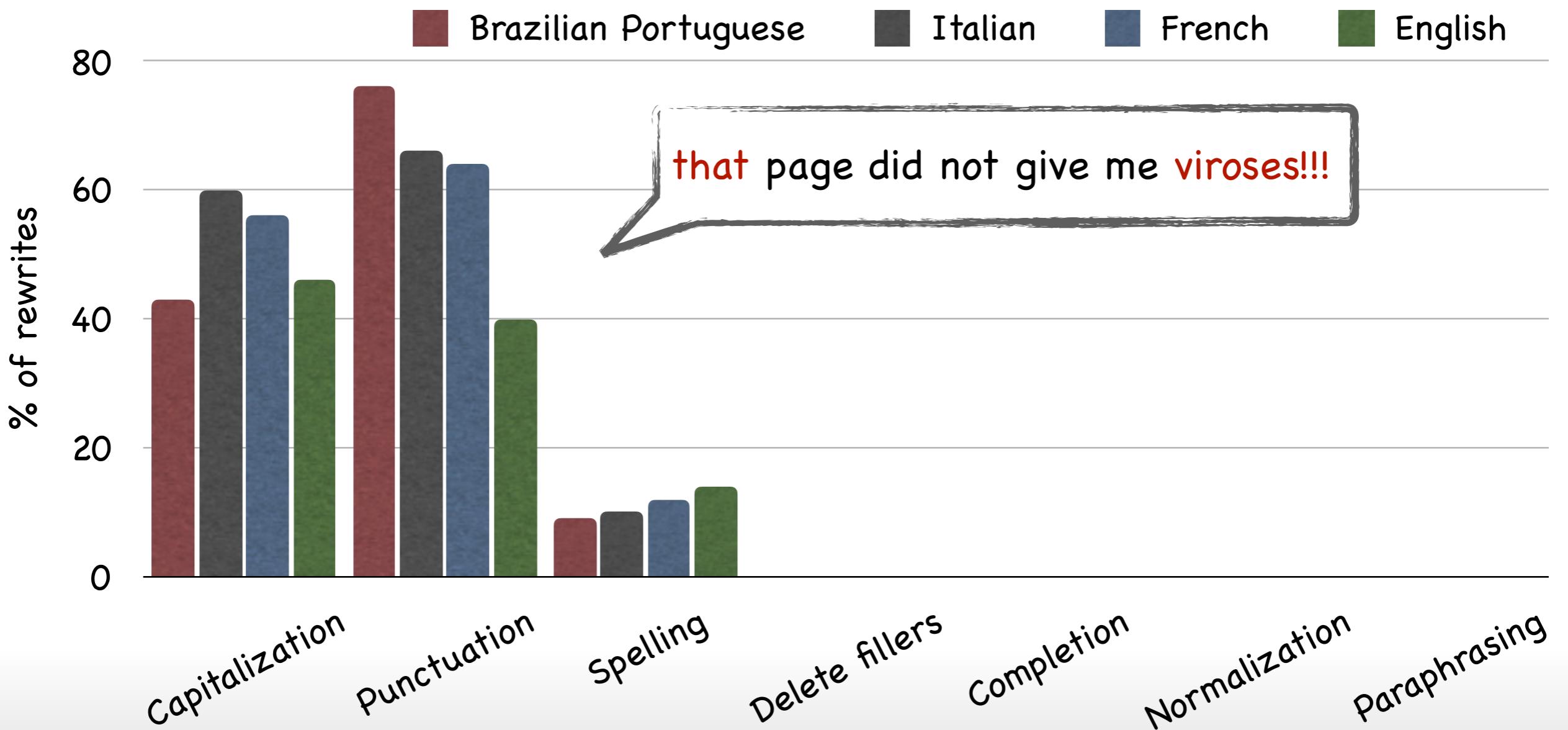
- QC1: Location Restrictions
- QC2: Qualification test
- QC3: Filtering by pilot study



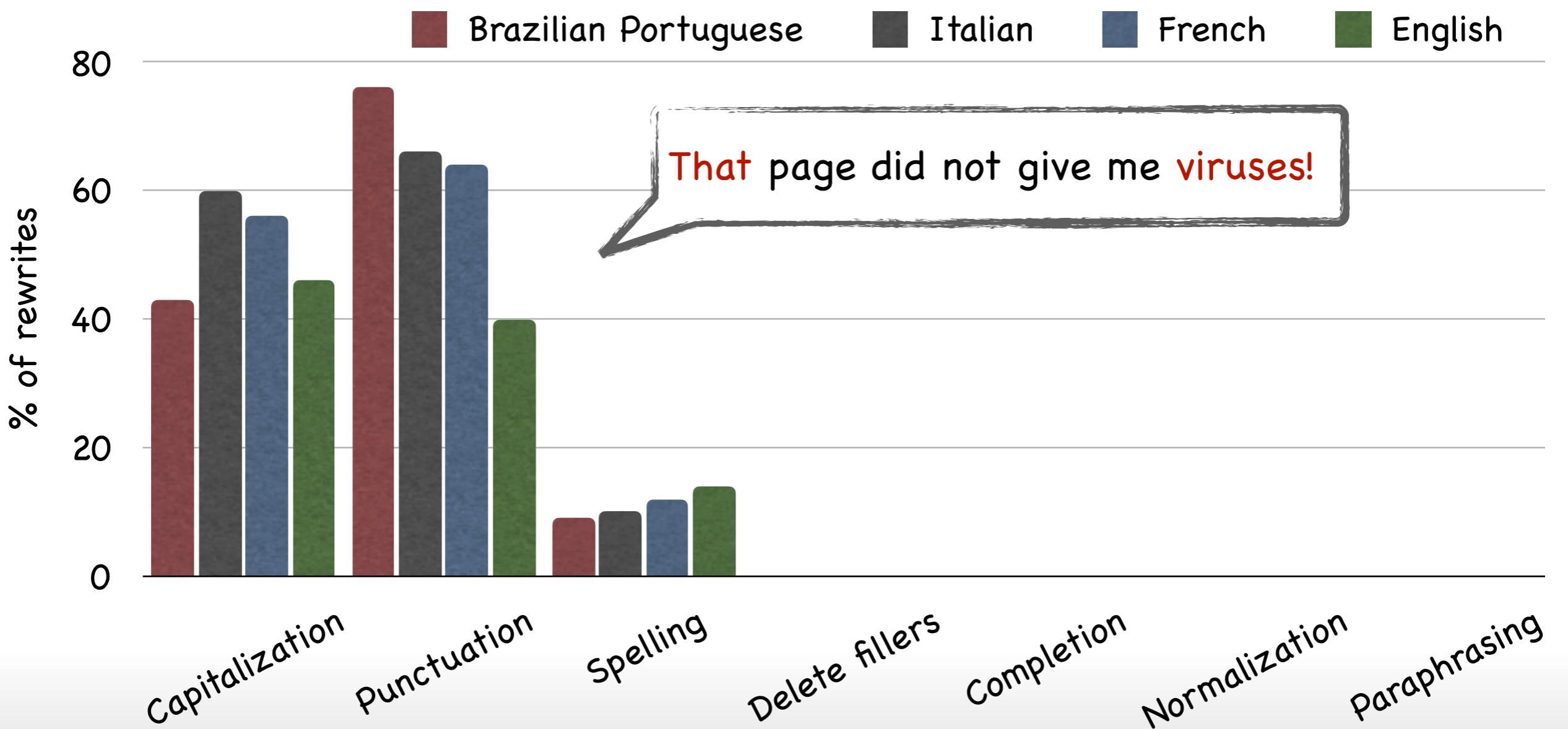
Formal edit operations across languages



Formal edit operations across languages

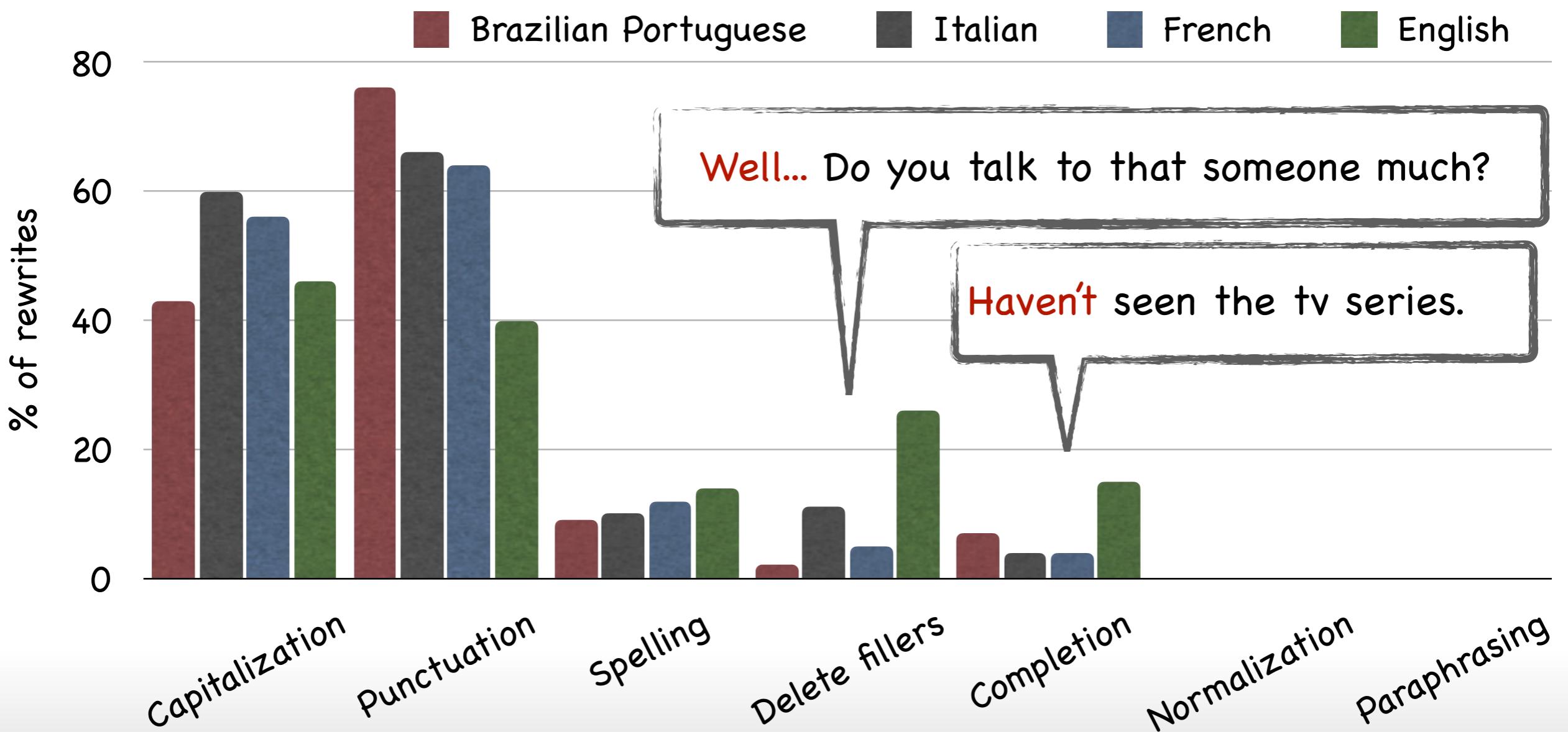


Formal edit operations across languages

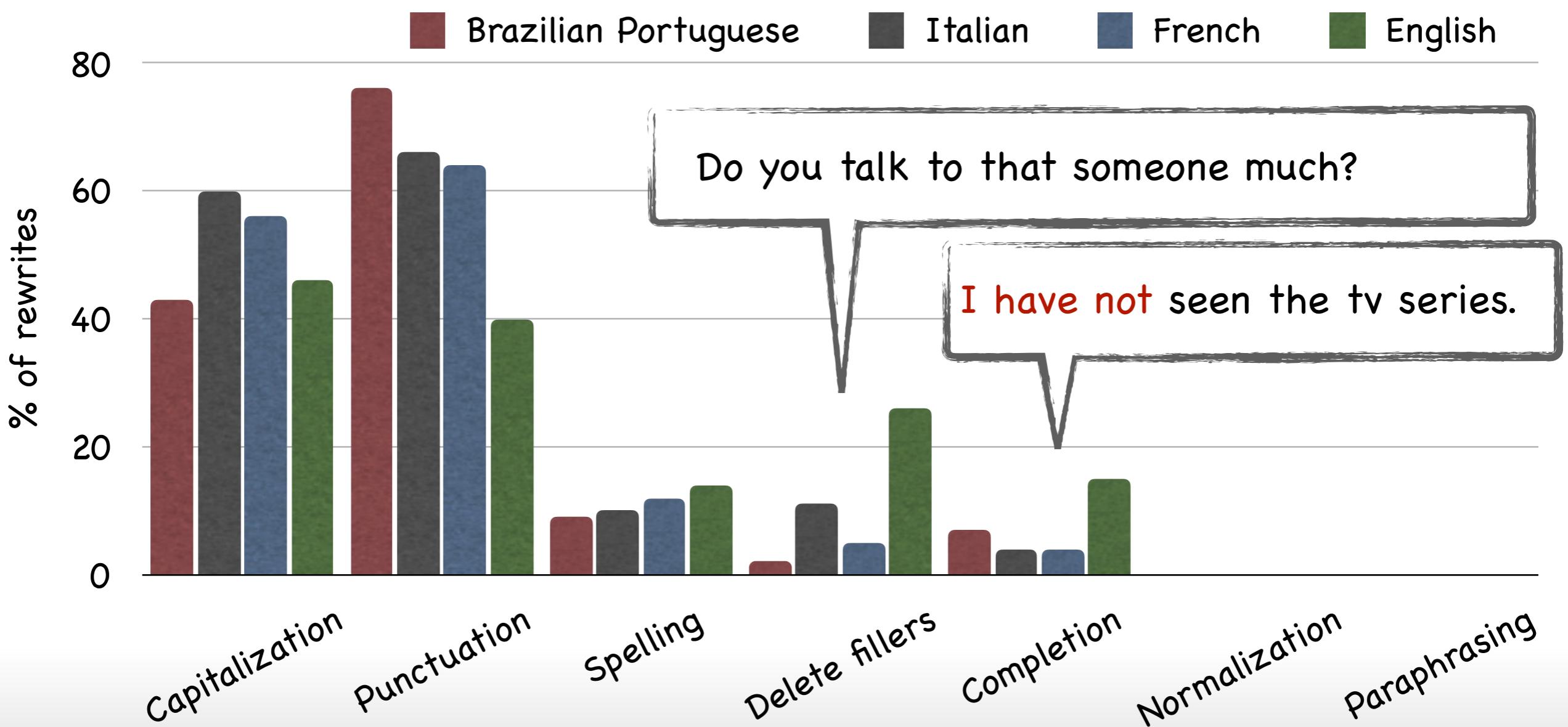


Consistent trends: edits cover the "noisy-text" sense of formality

Formal edit operations across languages

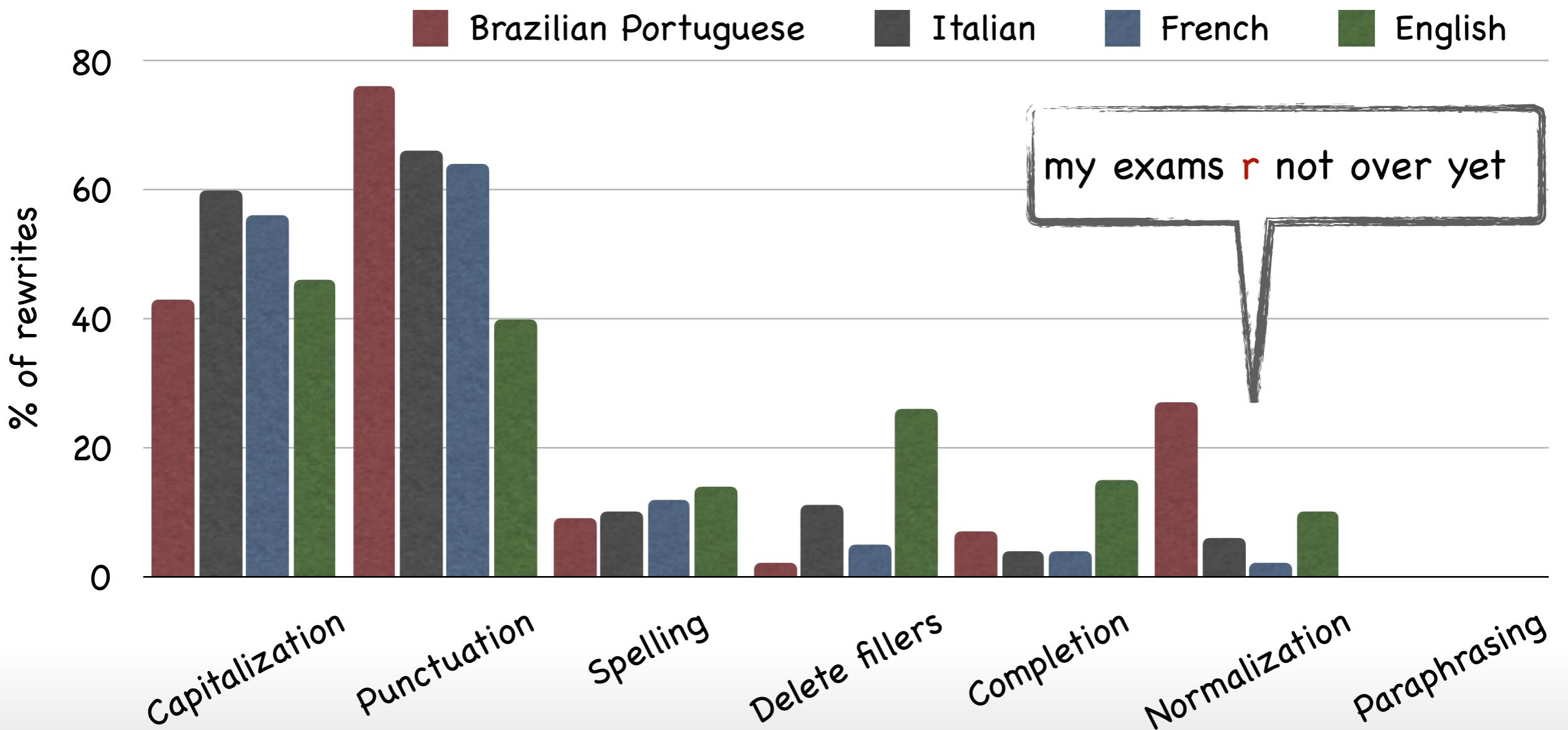


Formal edit operations across languages



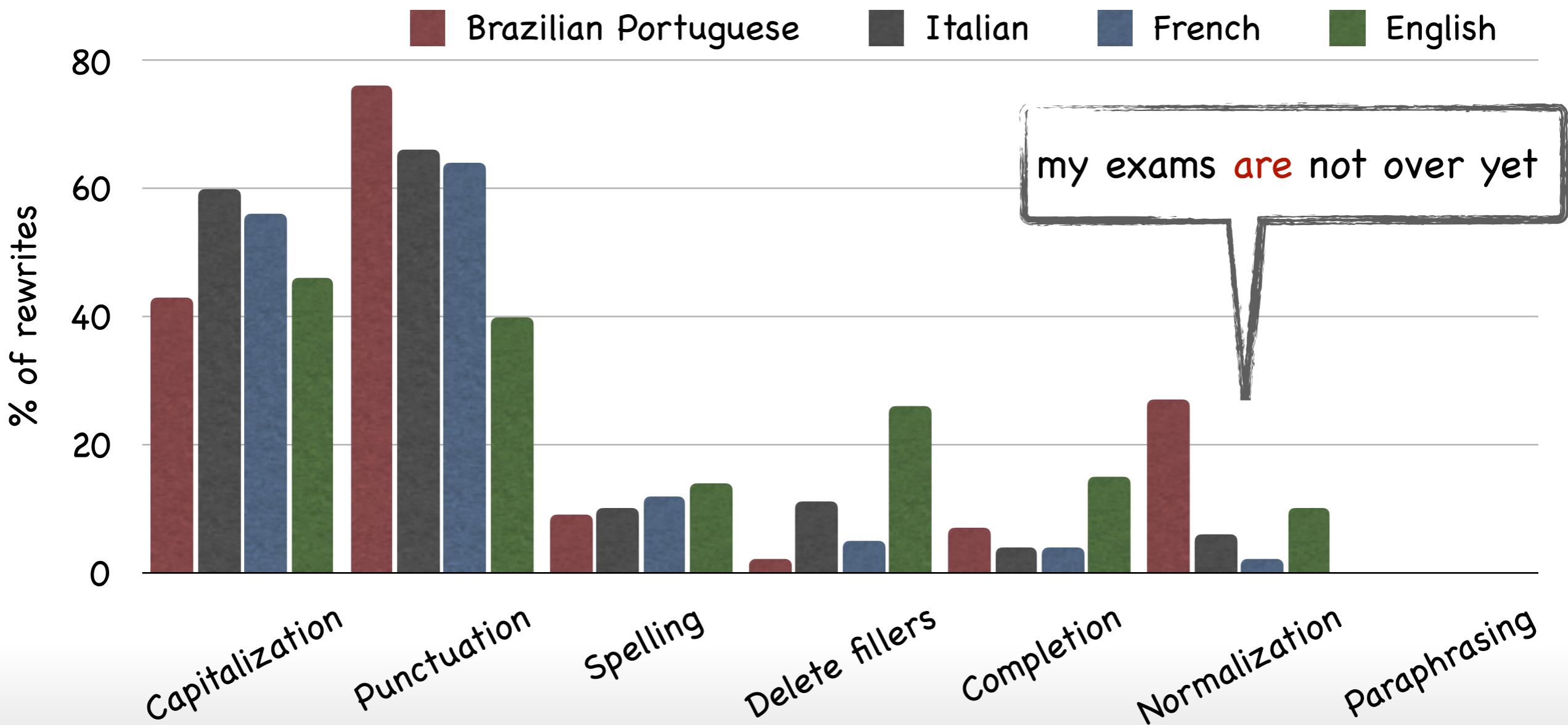
Delete fillers and completion-based edits are more frequent for EN

Formal edit operations across languages



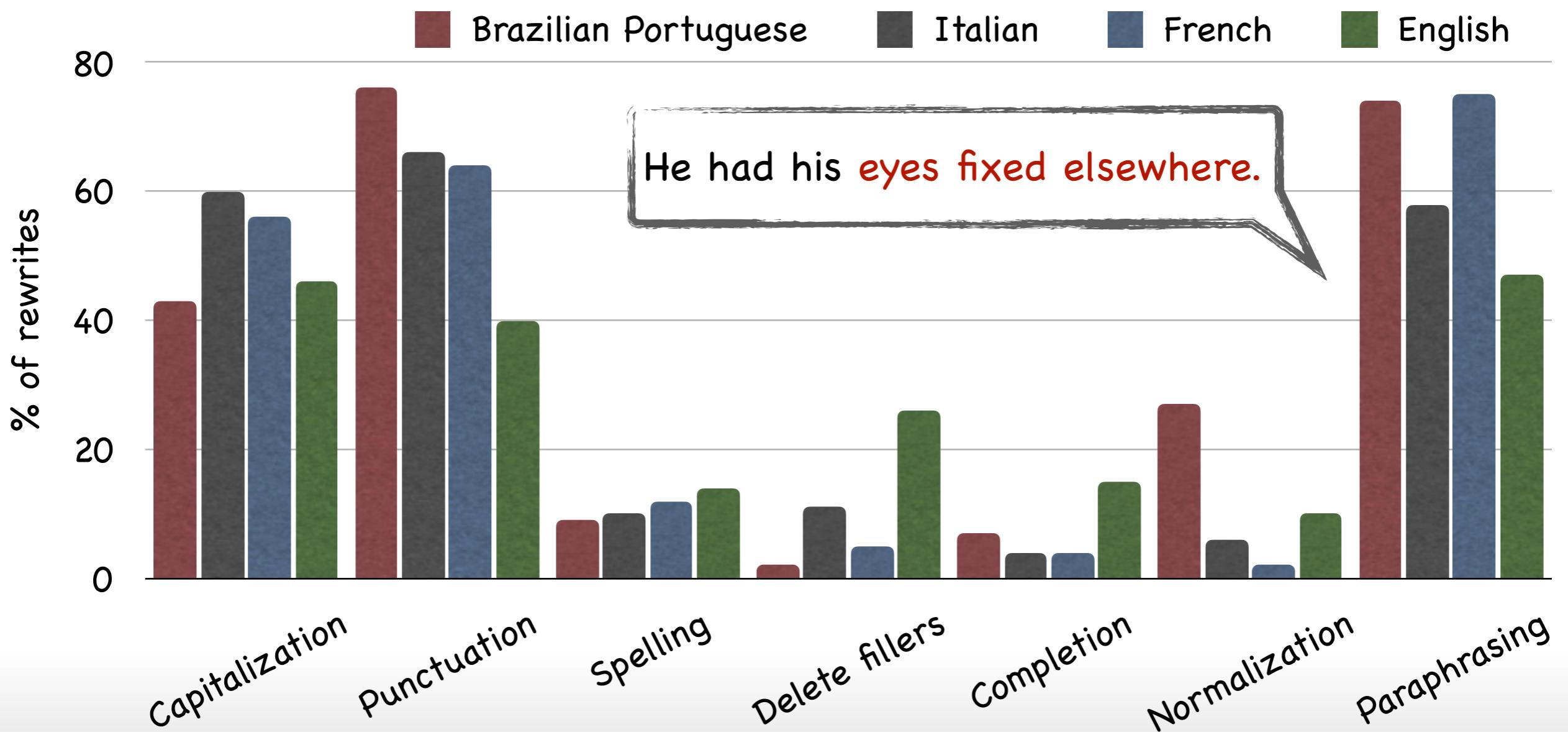
Normalization-based edits are more frequent for BR-PT

Formal edit operations across languages

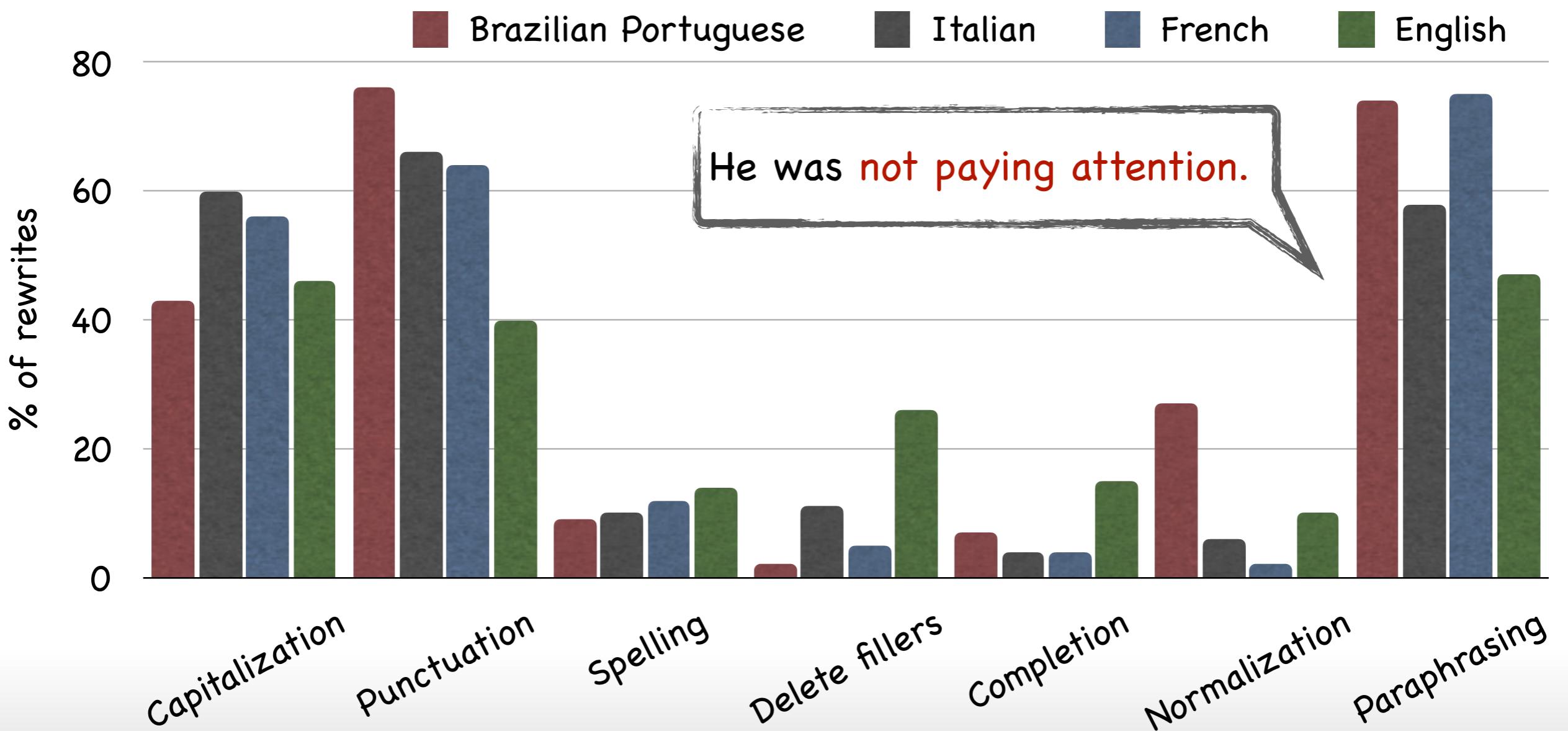


Normalization-based edits are more frequent for BR-PT

Formal edit operations across languages

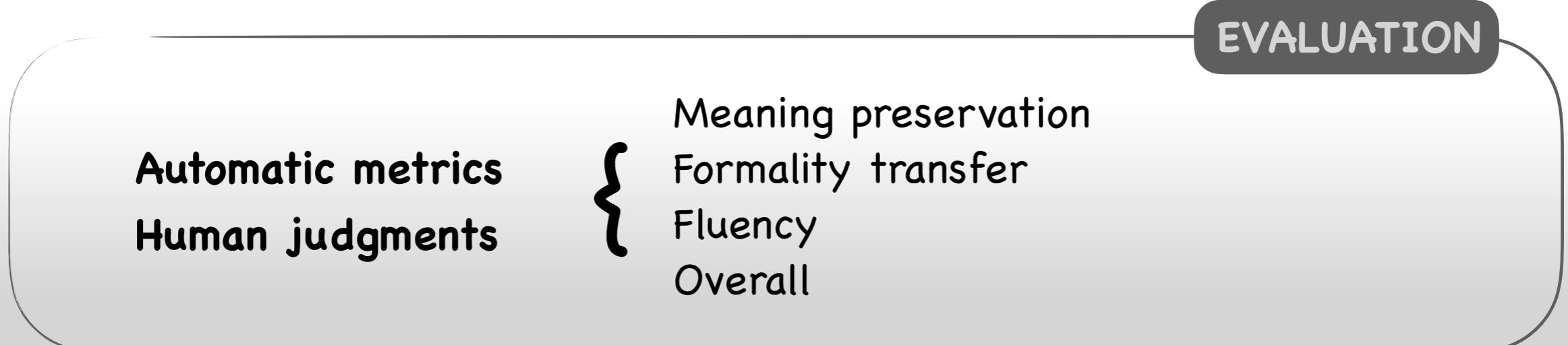
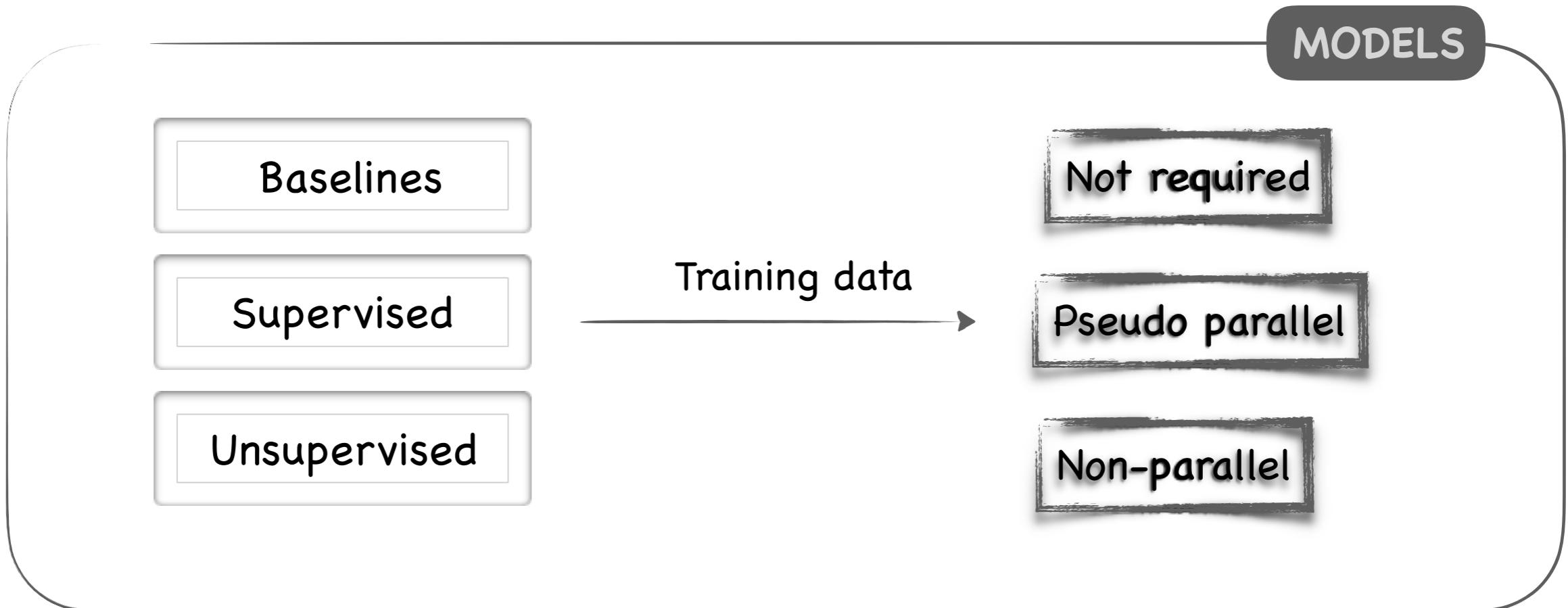


Formal edit operations across languages



Paraphrase-based edits are more frequent for non-En

Benchmarking multilingual FoST: Overview



Benchmarking multilingual FoST: Models

Simple Baselines

Copy

Rule-based

Round-trip NMT

Training data: Not required

Benchmarking multilingual FoST: Models

Simple Baselines

Copy

Rule-based

Round-trip NMT

INPUT

n preciso pedir pois sei q ela vai vir atras!!

No transformations

OUTPUT

n preciso pedir pois sei q ela vai vir atras!!

Training data: Not required

Benchmarking multilingual FoST: Models

Simple Baselines

Copy

Rule-based

Round-trip NMT

INPUT

n preciso pedir pois sei q ela vai vir atras!!

Hand-crafted transformations

Fix casing
Normalize punctuation
Expand contractions
etc.

OUTPUT

não preciso pedir pois sei **que** ela vai vir atras!

Training data: Not required

Benchmarking multilingual FoST: Models

Simple Baselines

Copy

Rule-based

Round-trip NMT

INPUT

n preciso pedir pois sei q ela vai vir atras!!

Pivot to EN and back-translate

AWS translation service
Formalization effect

OUTPUT

Não preciso perguntar porque sei que ela virá atrás de mim!

Training data: Not required

Benchmarking multilingual FoST: Models

NMT-based approaches

Translate-Train Tag

Multi-task Tag

Backtranslate

Niu et al; 2018

Sennrich et al; 2016

Training data: Pseudo-parallel

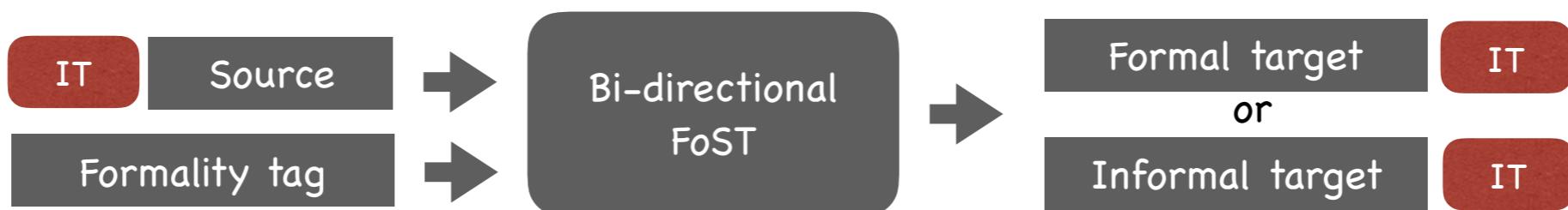
Benchmarking multilingual FoST: Models

NMT-based approaches

Translate-Train Tag

Multi-task Tag

Backtranslate



Side constraints

Training data: Pseudo-parallel

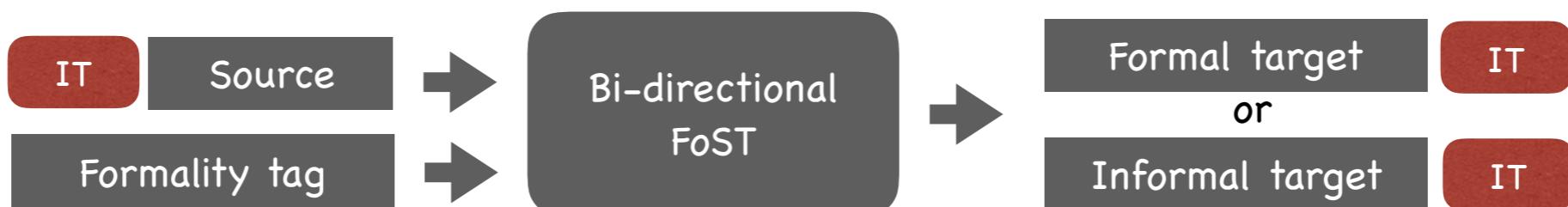
Benchmarking multilingual FoST: Models

NMT-based approaches

Translate-Train Tag

Multi-task Tag

Backtranslate



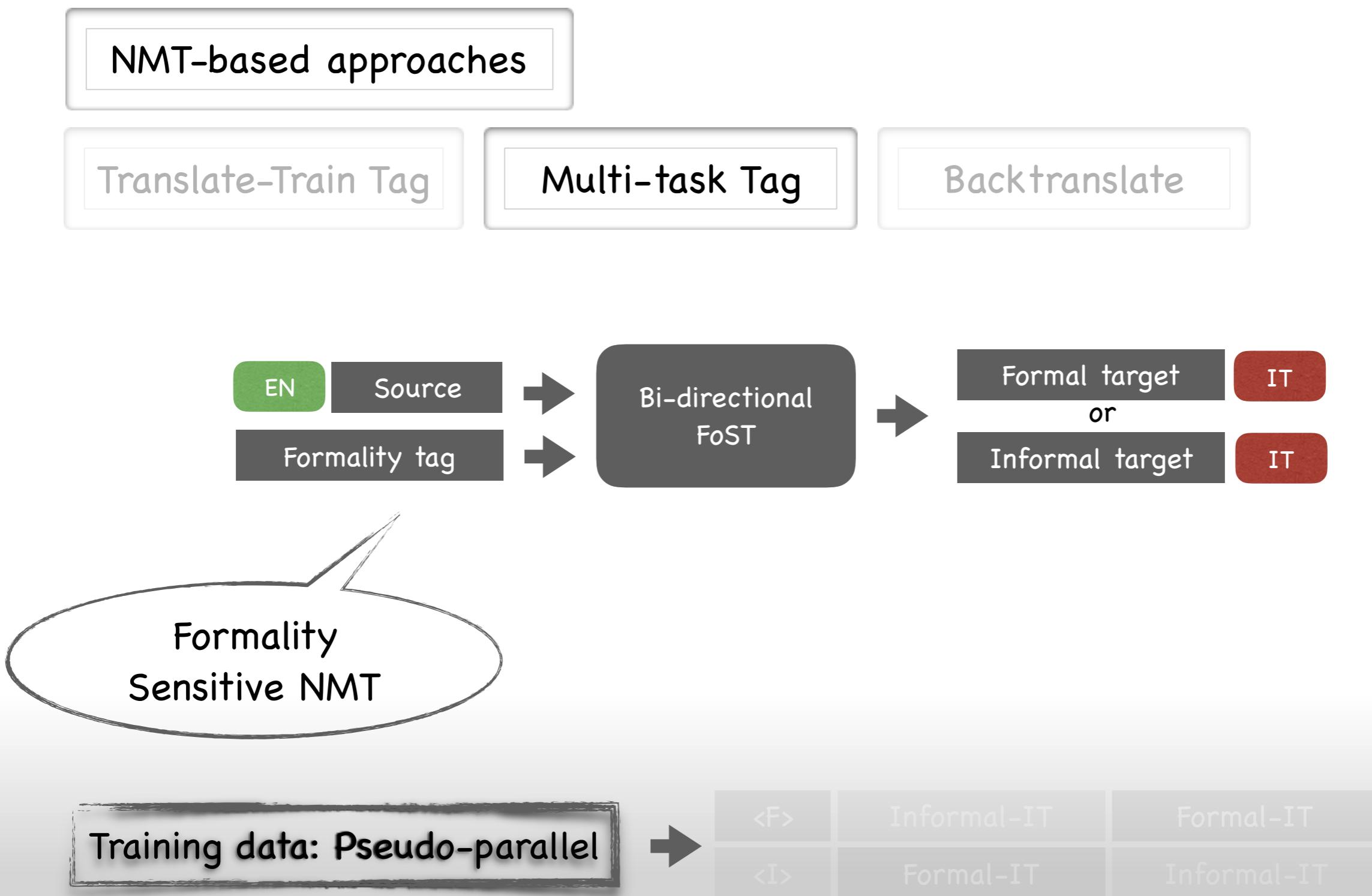
Side constraints

Translate-Train GYAFC

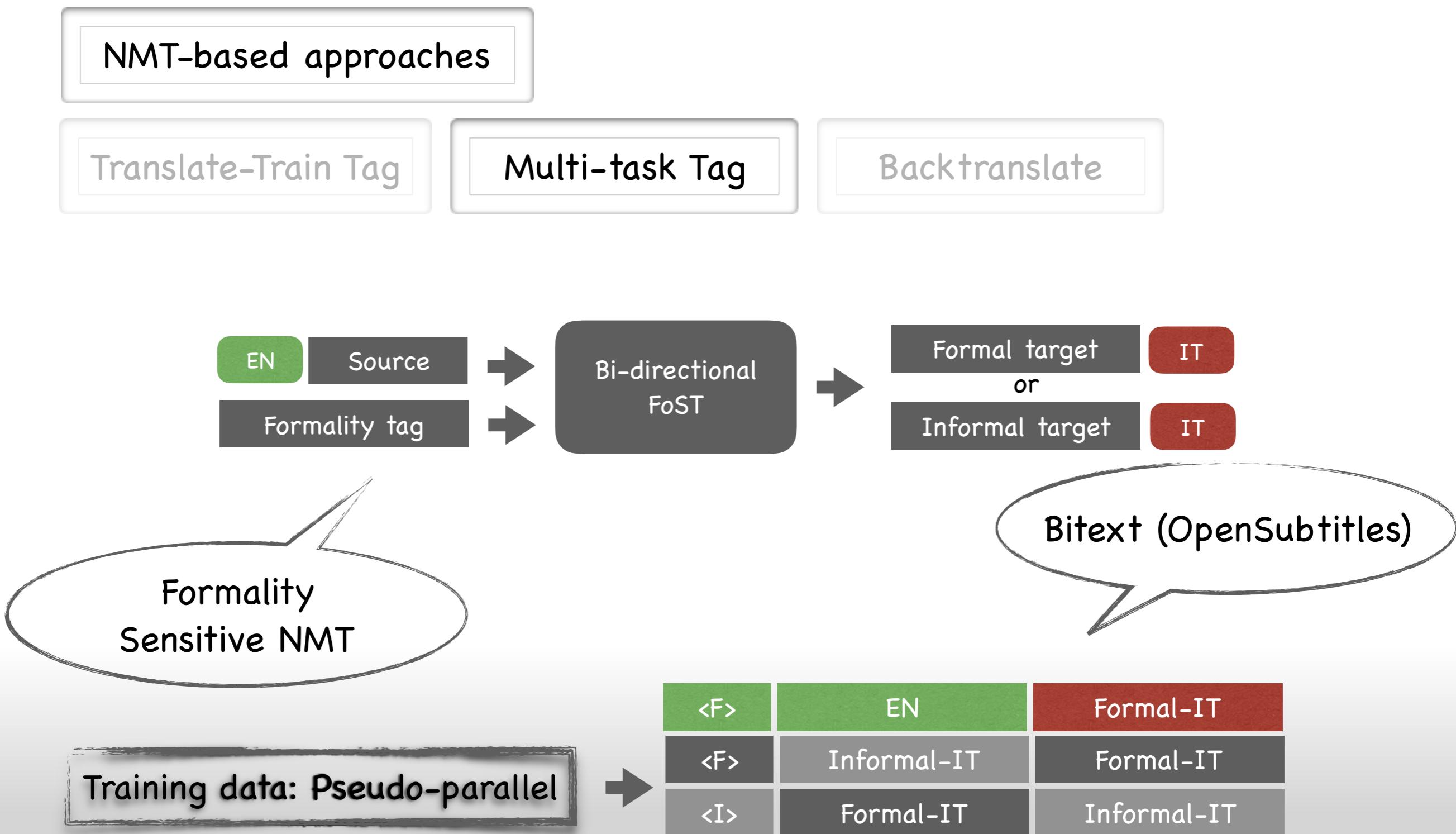
Training data: Pseudo-parallel

<F>	Informal-IT	Formal-IT
<I>	Formal-IT	Informal-IT

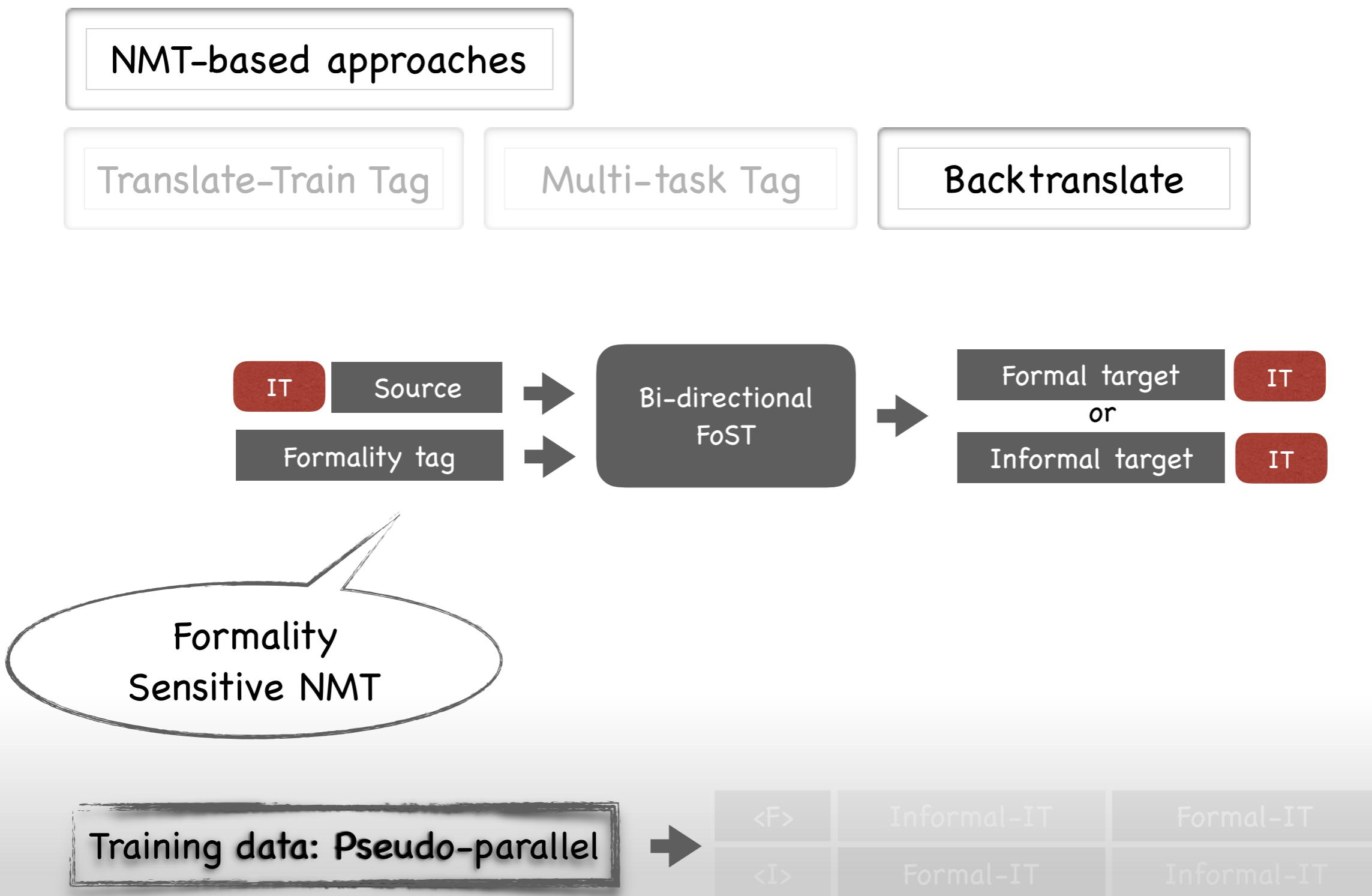
Benchmarking multilingual FoST: Models



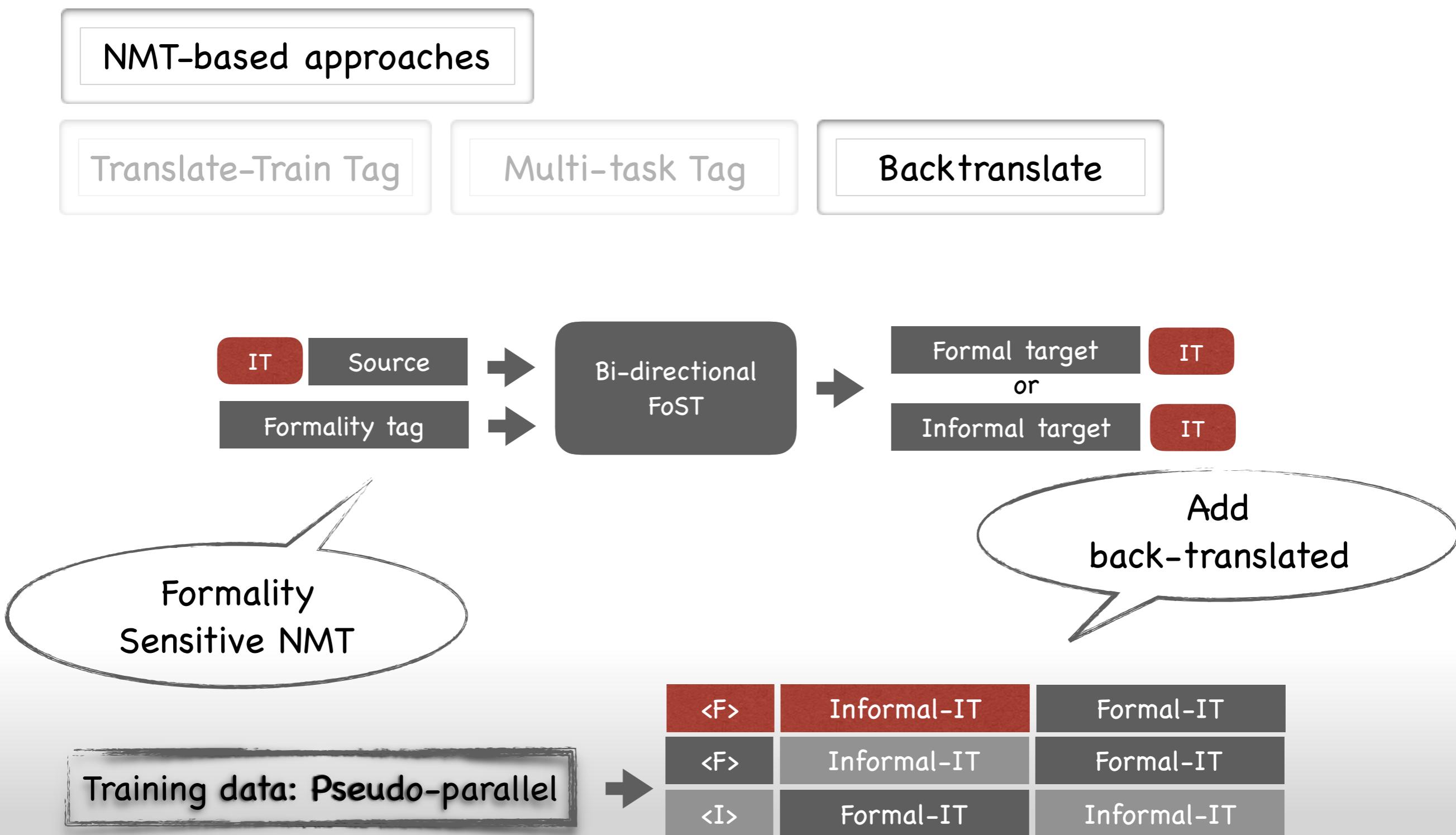
Benchmarking multilingual FoST: Models



Benchmarking multilingual FoST: Models

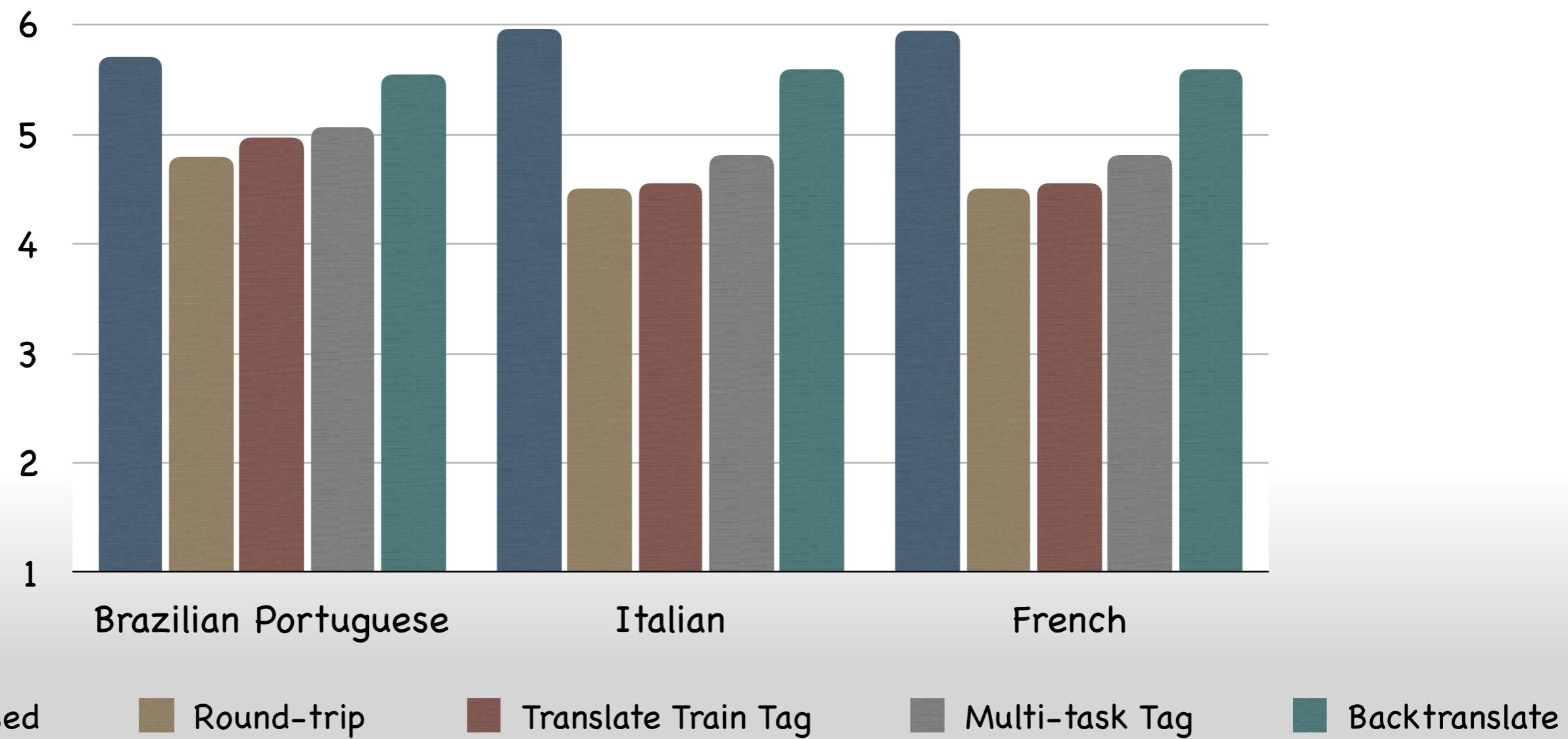


Benchmarking multilingual FoST: Models



Benchmarking multilingual FoST: Evaluation

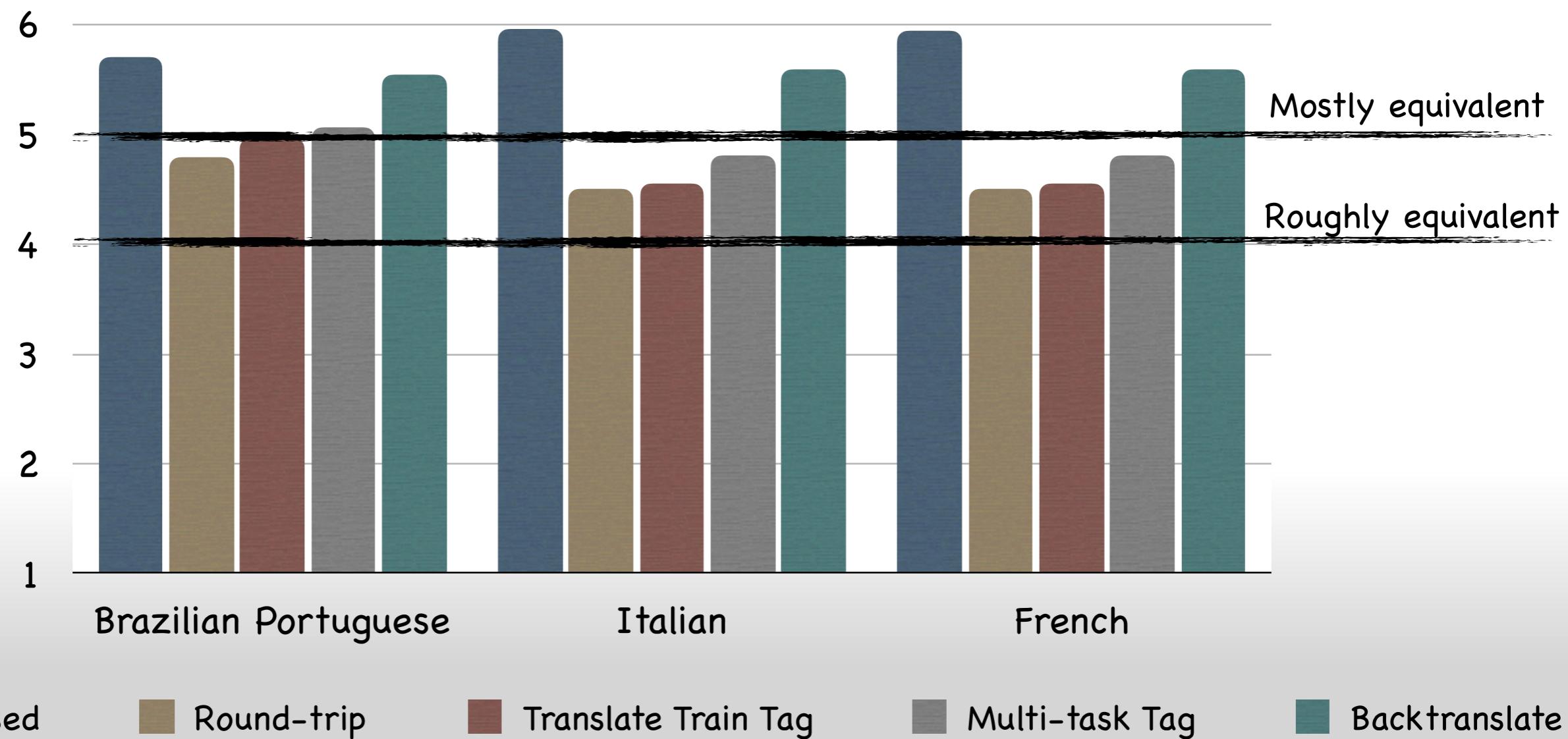
Meaning Preservation



Benchmarking multilingual FoST: Evaluation

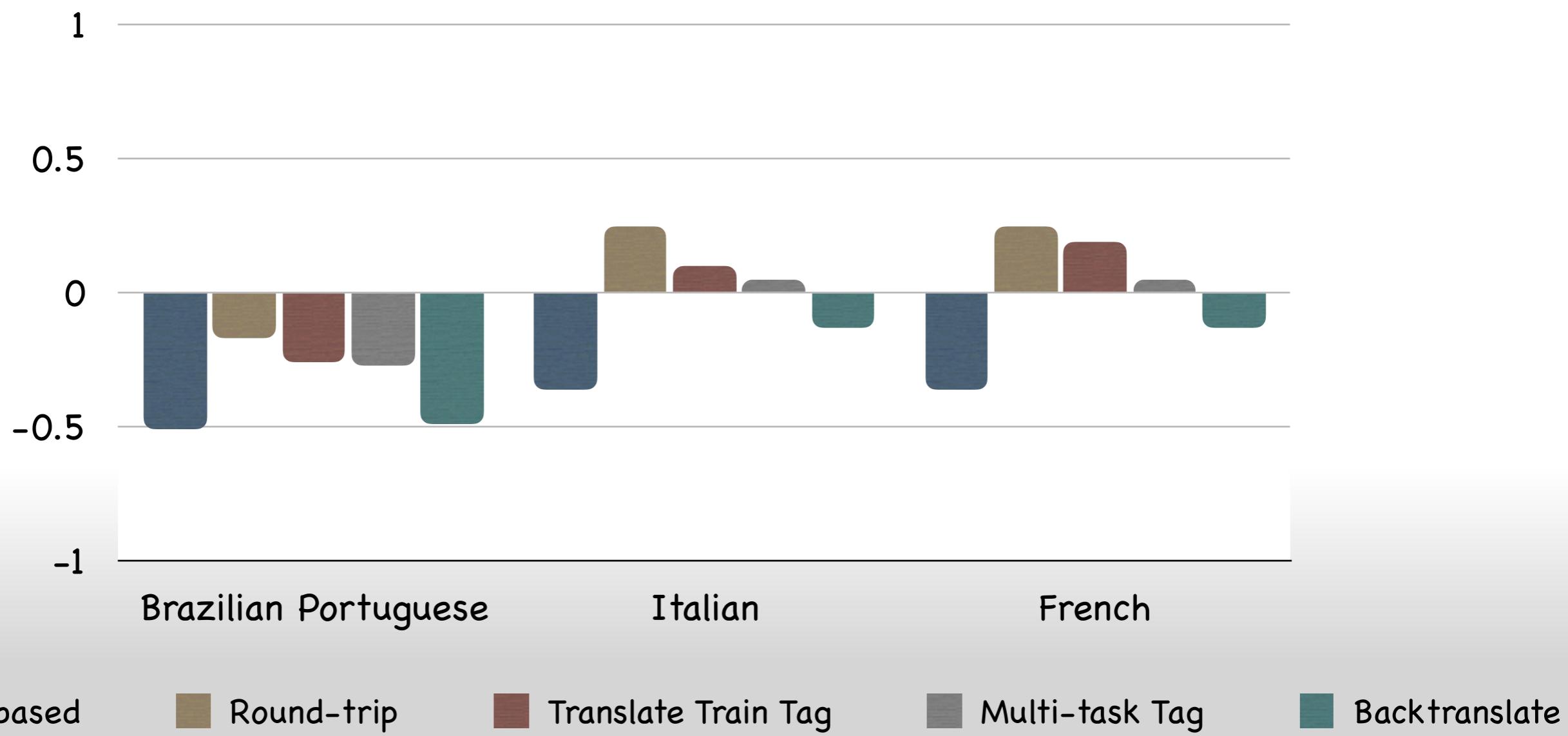
Meaning Preservation

System outputs are meaning preserving on average



Benchmarking multilingual FoST: Evaluation

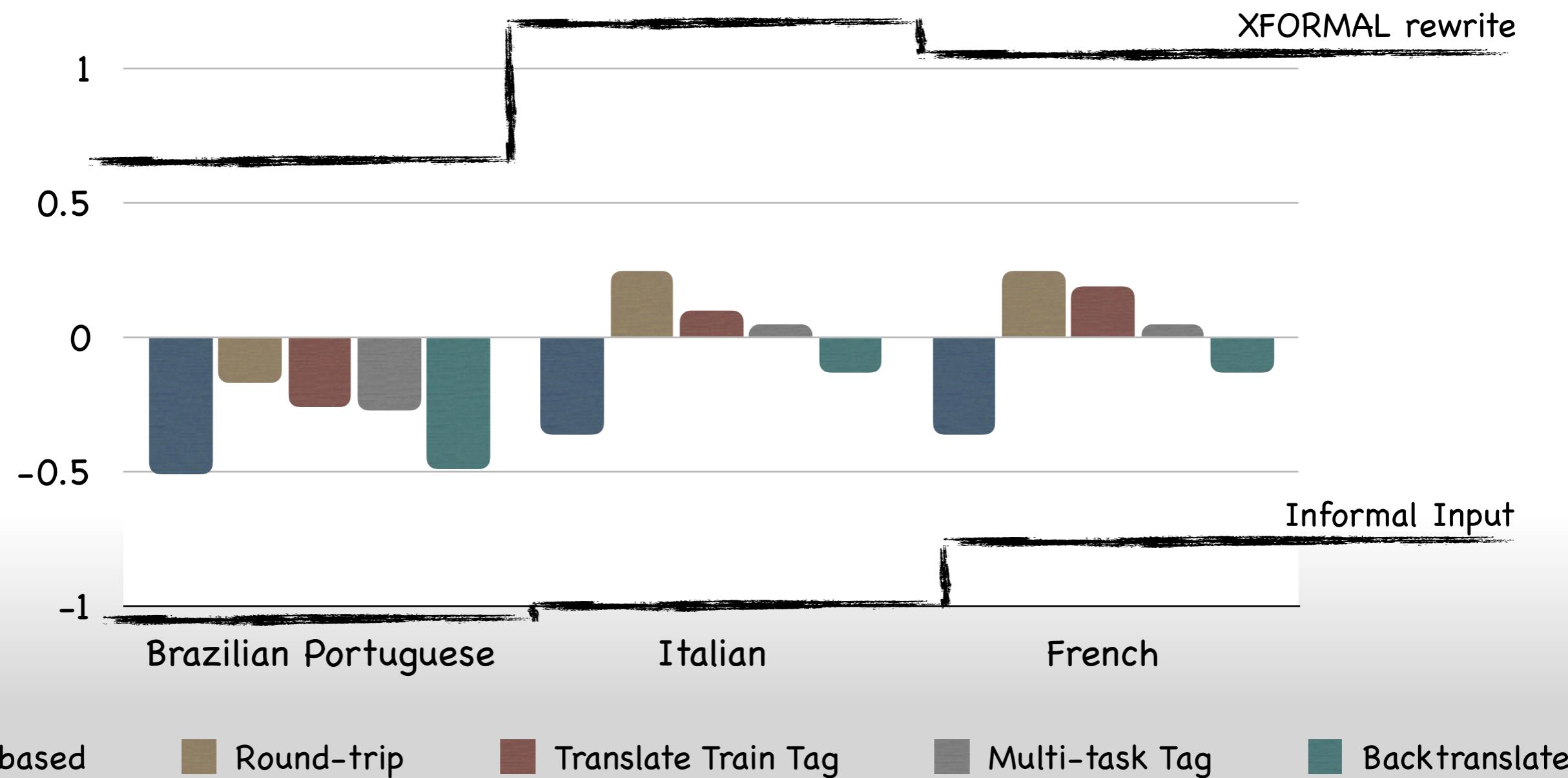
Formality Transfer



Benchmarking multilingual FoST: Evaluation

Formality Transfer

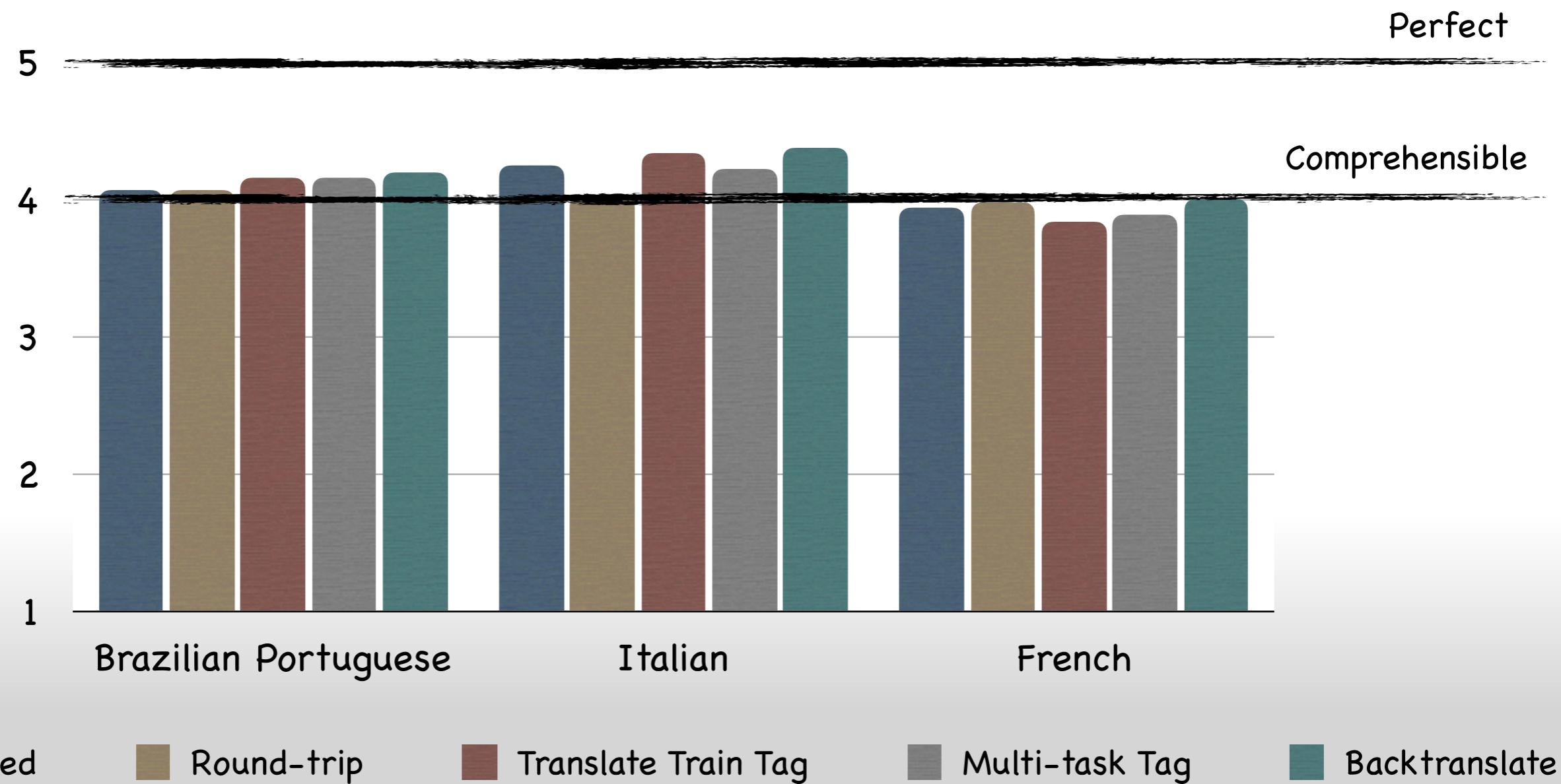
System outputs are concentrated around neutral formality levels



Benchmarking multilingual FoST: Evaluation

Fluency

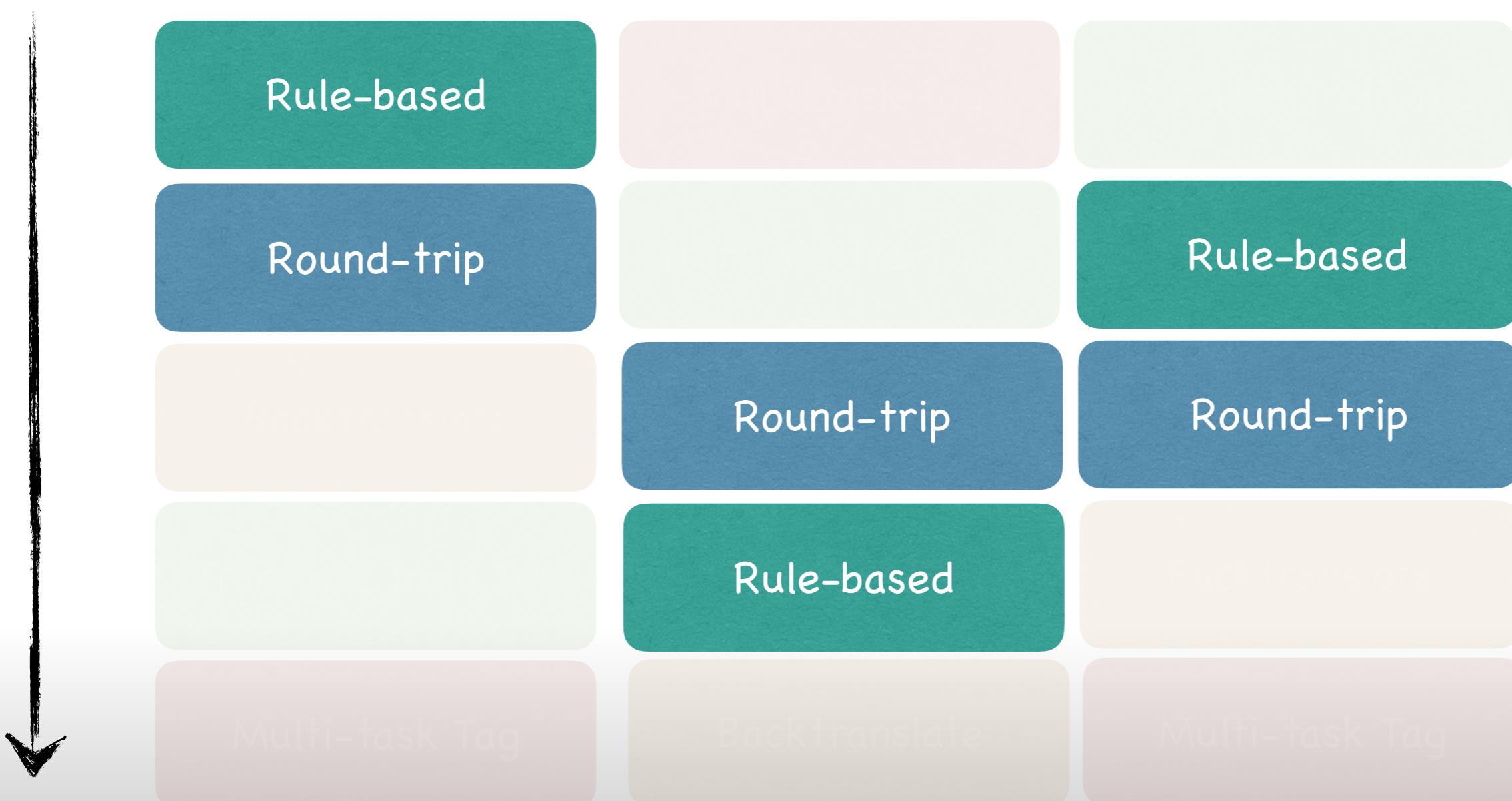
System outputs are comprehensible on average



Benchmarking multilingual FoST: Evaluation

Worst	Brazilian Portuguese	Italian	French
	Rule-based	Multi-task Tag	Translate Train Tag
	Round-trip	Translate Train Tag	Rule-based
	Backtranslate	Round-trip	Round-trip
	Translate Train Tag	Rule-based	Backtranslate
	Multi-task Tag	Backtranslate	Multi-task Tag
Best	Overall Ranking		

Benchmarking multilingual FoST: Evaluation



Simple baselines perform comparable to more advanced FoST models

Summary

- XFORMAL

- Evaluation dataset of informal-formal pairs in FR, IT, BR-PT
 - MTurk & multiple levels of quality control

- Benchmarking of FoST

- most systems perform conservative edits on the informal input
 - simple baselines perform comparable to advanced models

- Future work should look at...

- models that do not heavily rely on supervised data
 - automatic evaluation methods that generalize beyond English

Summary

- XFORMAL

- Evaluation dataset of informal-formal pairs in FR, IT, BR-PT
 - MTurk & multiple levels of quality control

- Benchmarking of FoST

- most systems perform conservative edits on the informal input
 - simple baselines perform comparable to advanced models

- Future work should look at...

- models that do not heavily rely on supervised data
 - automatic evaluation methods that generalize beyond English