Multilingual BERT has an accent

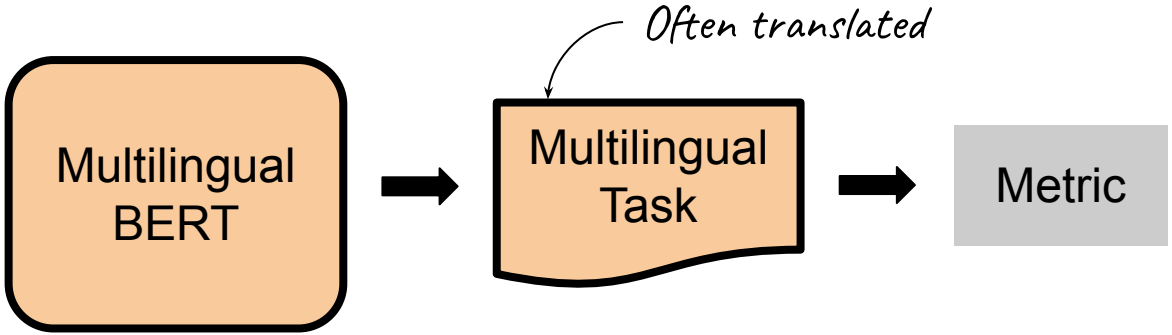
Evaluating English influences on fluency in multilingual models

Isabel Papadimitriou*, Kezia Lopez*, Dan Jurafsky

Stanford NLP

Multilingual Models

- Impressive, cross-language results!



- This paper: **fluency evaluation,** the effect of English syntactic structure on co-trained languages.

Main idea: focus on constructions that are in free variation

Pick a construction in a language which has two forms: one which is parallel to English, one which is not

Spanish: Pronoun drop

(1)
Entonces **ella toma** la bandera de la revolución
So **she take.3SG** the flag of DET revolution

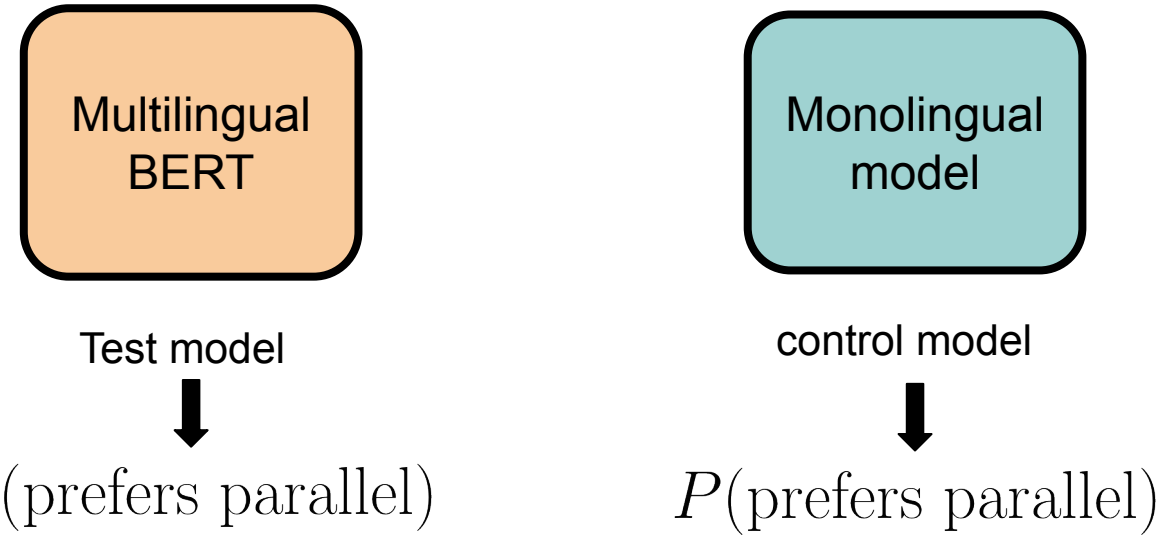
(2)
Escribió numerosas obras de historia
write.3SG.PST numerous works of history

Greek: Subject-Verb order

(3)
Ο πρώτος **αγώνας έληξε** με σκορ 3:2
The first **match** **finished** with score 3:2

(4)
Το σκορ του αγώνα **άνοιξε** ο **Γουέν** Ρούνι
The score of match **opened** DET **Wayne** Rooney

Does mBERT prefer the English-like construction?

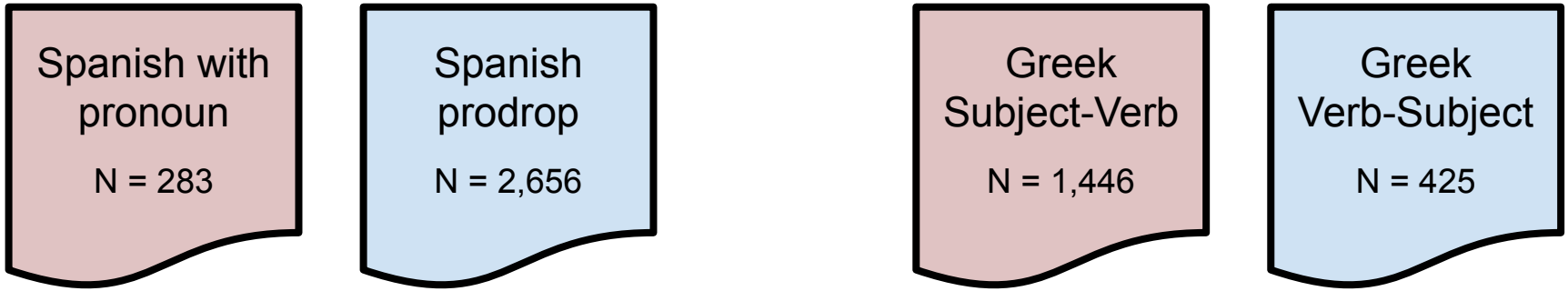


Method:

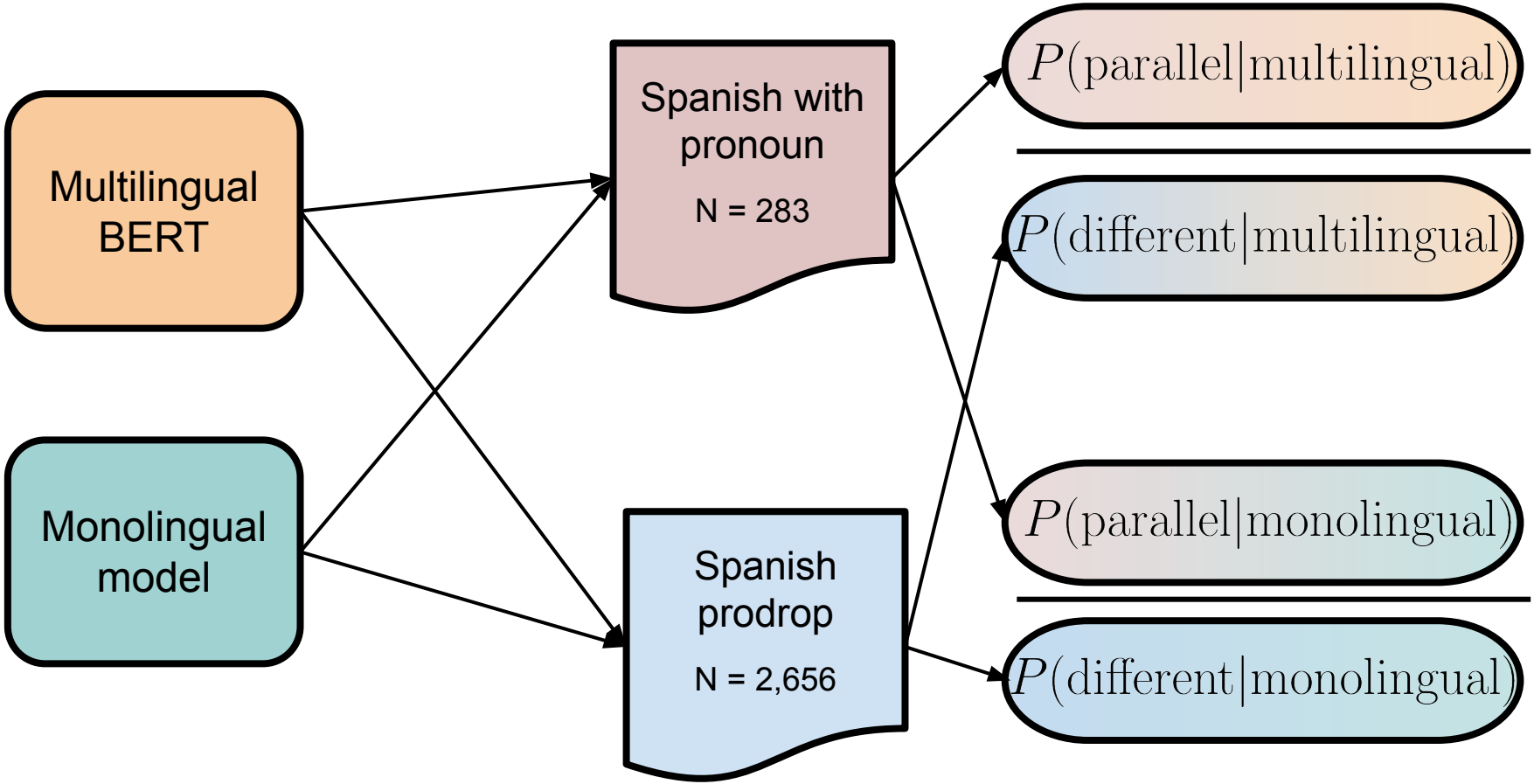
Getting an approximation for $P(\text{prefers parallel})$

Data

- Pick out sentences from UD to make the **parallel** and **different** corpora



- Important that we don't construct data – we're looking for fluency


Multilingual BERT
Monolingual model
Spanish with pronoun
N = 283
Spanish prodrop
N = 2,656
$P(\text{parallel}|\text{multilingual})$
$P(\text{different}|\text{multilingual})$
$P(\text{parallel}|\text{monolingual})$
$P(\text{different}|\text{monolingual})$

$$P(\text{prefers parallel}|\text{model}) = \frac{P(\text{parallel}|\text{model})}{P(\text{different}|\text{model})}$$

How to approximate probability of construction?

- Pick one word in each construction to represent it

Spanish: the verb

Entonces **ella** **toma** la bandera de la revolución

Escribió numerosas obras de historia

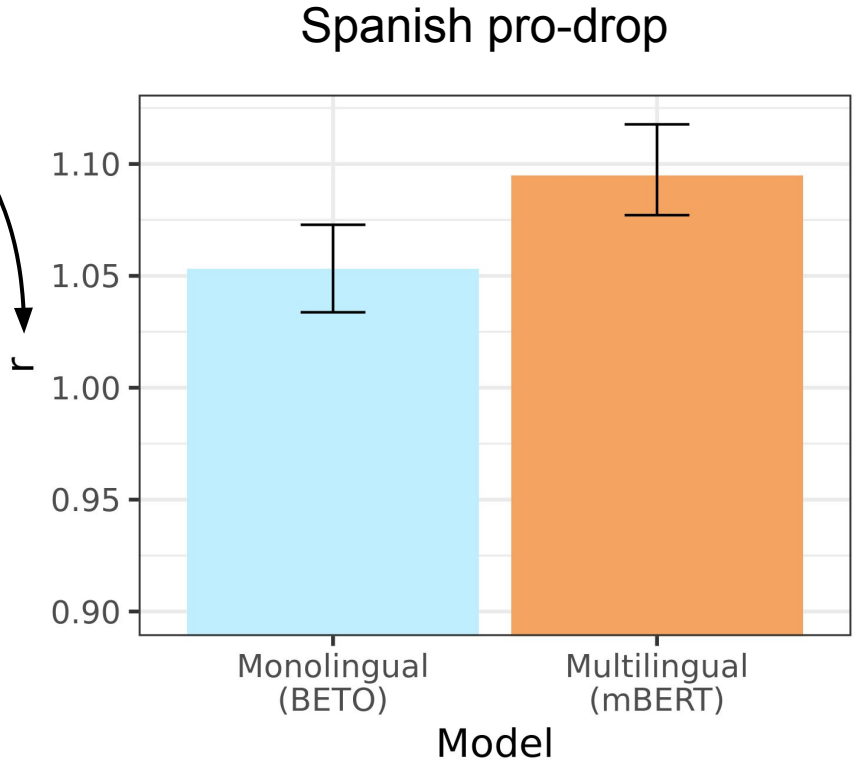
Greek: first word of subject/verb

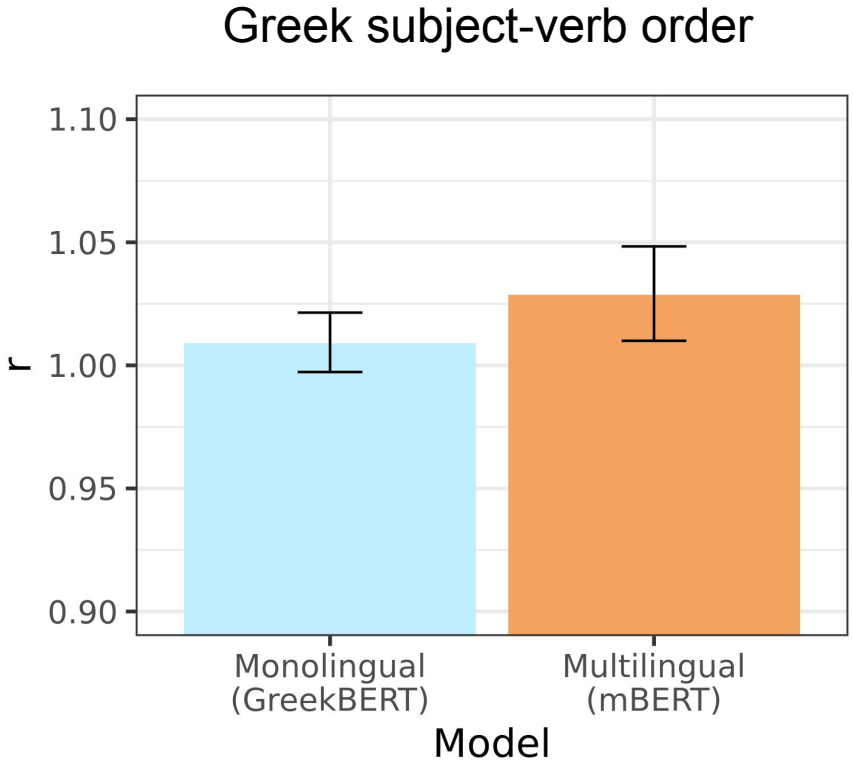
Ο πρώτος **αγώνας** **έληξε** με σκορ 3:2

Στις 3_Σεπτεμβρίου **ξέσπασε** **επανάσταση**

Results

$$P(\text{prefers parallel}|\text{model}) = \frac{P(\text{parallel}|\text{model})}{P(\text{different}|\text{model})}$$





Fluency evaluation

- The effect probably won't influence downstream, non-generation performance very much
- And there are so many other things that we're not looking at
  - Subtleties in lexical choice, grammatical choice, discourse…
- Entanglement of fluency with domain variation, translationese
- Future work: evaluations that don't use a monlingual model
  - Preserving stimuli **natruralness** in a **controlled** experiment

Thanks!