Part-of-Speech Annotation in 4 hours: How to spend your time

Dan Garrette, Jason Mielens, and Jason Baldridge

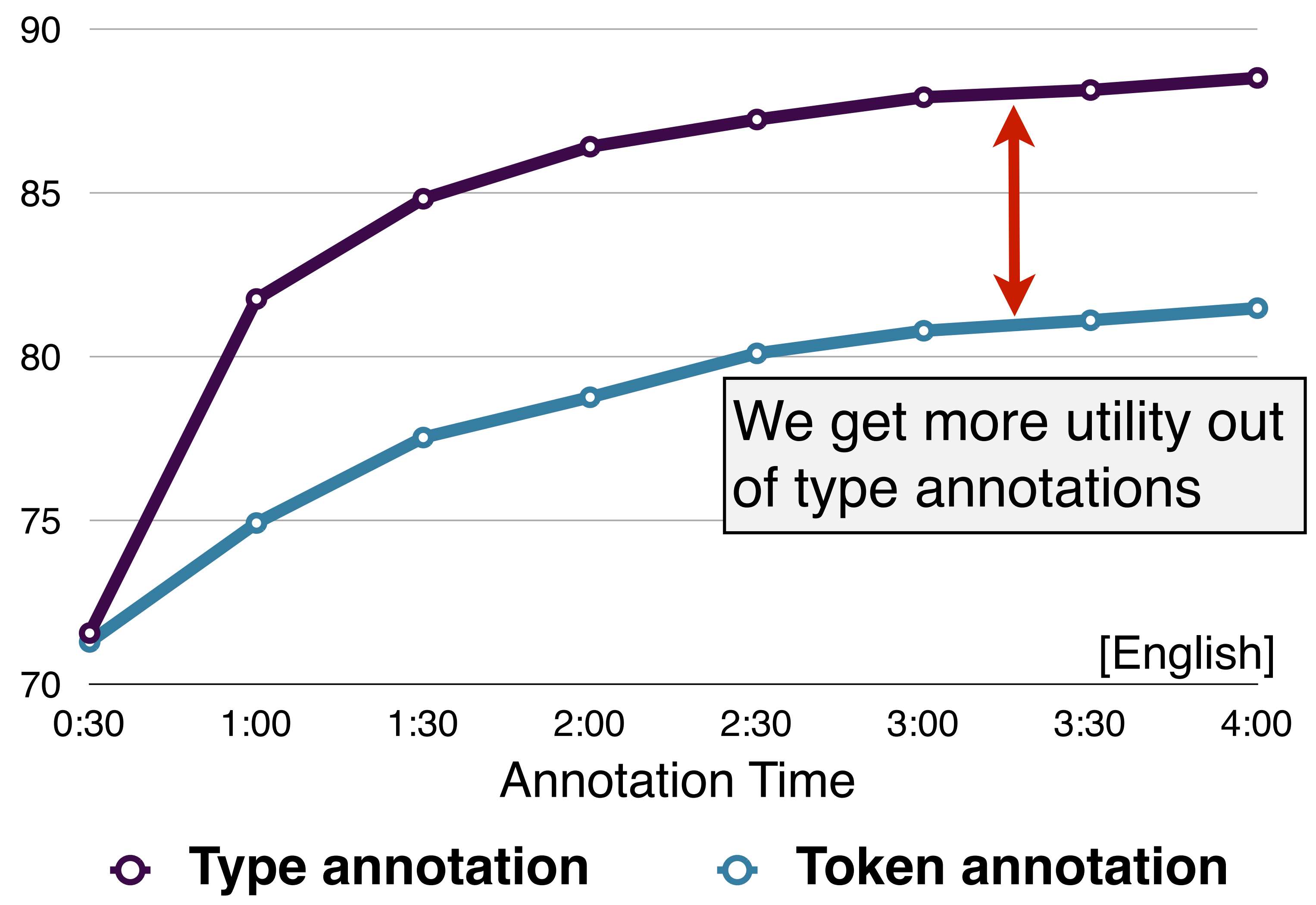
The University of Texas at Austin

For low-resource languages, we cannot expect much training data.

- Can we train with less data?
- What can we collect cheaply?
- Where to focus collection efforts?

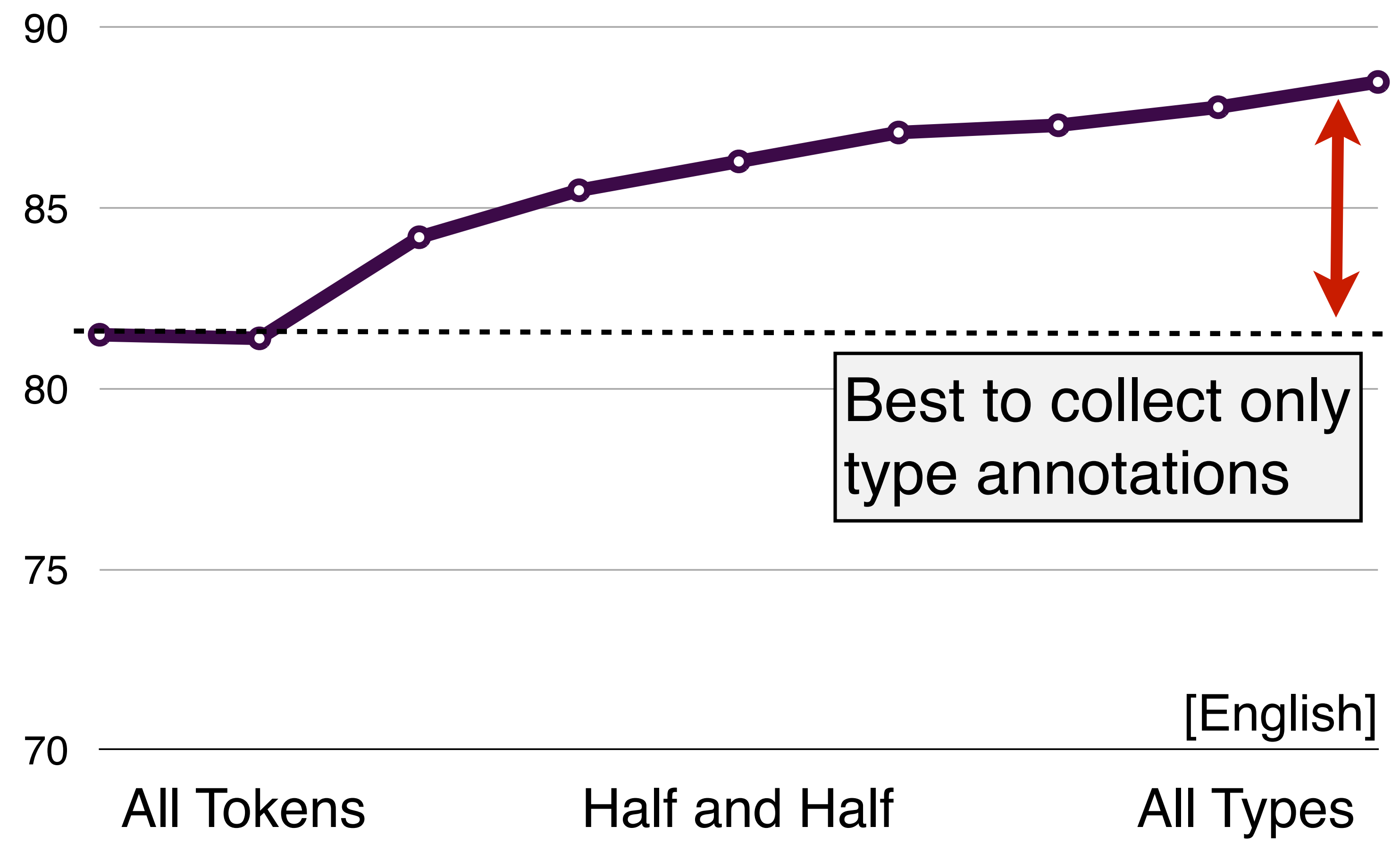
We conducted experiments collecting data from linguists in a fixed time.

Annotating: Types > Tokens



We get more utility out of type annotations

Mixing Type and Token Annotations



Best to collect only type annotations

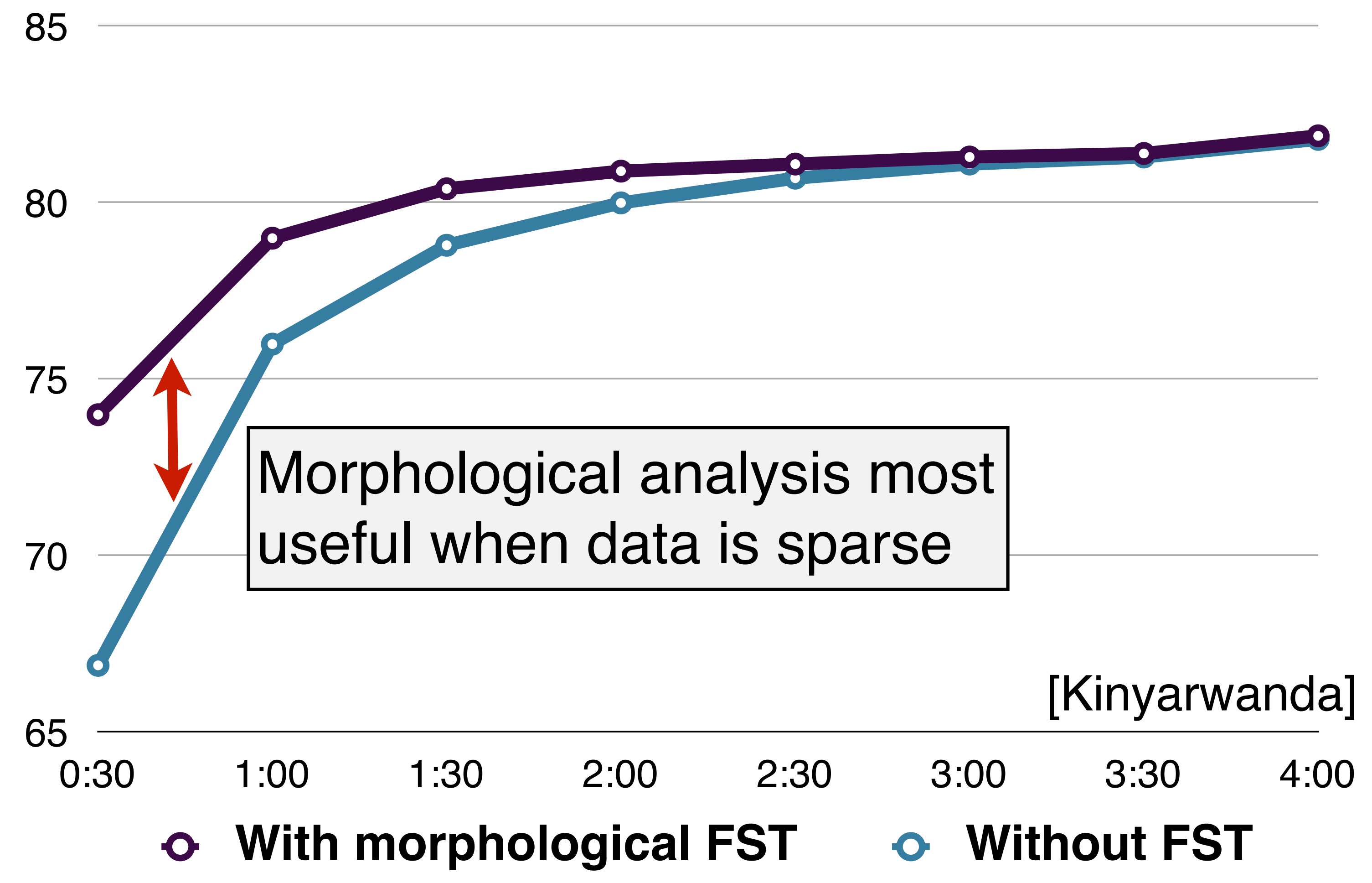
Our semi-supervised approach:

- Expand the initial annotations
- Remove expansion noise by finding a minimal model
- Train a MaxEnt Markov Model

Achieve **90%** accuracy after **4 hours**

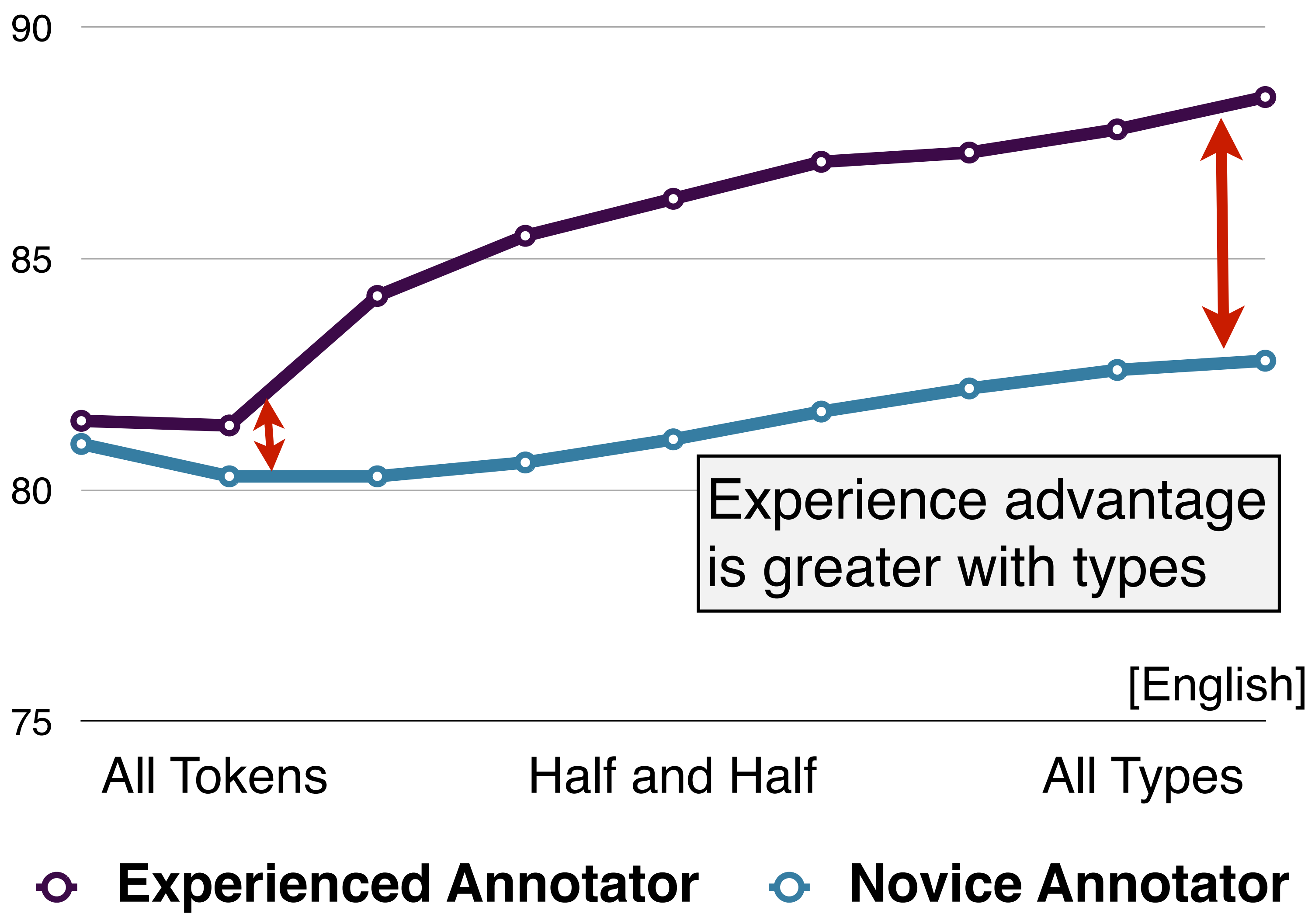
[Garrette & Baldridge, NAACL 2013]

Morphological analysis with an FST helps for Kinyarwanda



Morphological analysis most useful when data is sparse

Annotator Experience



Experience advantage is greater with types