

### 1. Preprocessing

1. Regex cleaning
2. Sentence Segmentation
3. Tokenization
4. Annotation

[0, Headingley is a suburb of  
Leeds. It is the location ...,

1, Leeds is the largest city in  
West Yorkshire. ...,

2, ...]

### 2. Annotated Wikipedia Dataset

Headingley	PLACE
is	O
a	O
suburb	O
of	O
Leeds	PLACE

### 3. Model and Evaluation

Embedding

Token Classification Model

SpaCy NER

Stanza NER

Evaluation

Test Subset

