You find a POS-tagged corpus of text in an unknown language. Can you parse this?

### Surface Cues to Structure
- Let’s extract interesting features of the whole corpus (“surface cues to structure”).
- Our universal parser sees these corpus features, along with the input sentence.
- The universal parser is trained end-to-end on diverse languages, with supervision from treebanks.
- Including treebanks for thousands of synthetic training languages. This helps.
- Our best method improved UAS and LAS on held-out test languages by an average of 5.6 percentage points over past work.

### The Typology Component
- Hand engineered features
  - How often do NOUNS tend to appear shortly before or after VERBs?
  - How often do ADJs tend to appear shortly before or after NOUNS?
  - How often do ADPs tend to appear shortly before or after NOUNS?

### What are we building?
- POS seq.
- POS corpus
- Unseen language
- Predicted arcs (dep. tree)

### How do we train?
- BiLSTM
- Treebanks
- Discard trees
- Treebanks

### Results (each bar stretches from labeled to unlabeled score)
- ~8000 (real + synthetic)
- 20 (real)

### Galactic Treebanks (Wang & Eisner 2016)
- More than 50,000 synthetic languages
- Resemble real languages, but not found on Earth
- Each has a corpus of dependency parses
- In the Universal Dependencies format
- Vertices are words labeled with POS tags
- Edges are labeled syntactic relationships
- Provide train/dev/test splits, alignments, tools