[TACL] Surface Statistics of an Unknown Language Indicate How to Parse It

Dingquan Wang and Jason Eisner
{wdd,eisner}@jhu.edu

The Center For Language and Speech Processing at the Johns Hopkins University

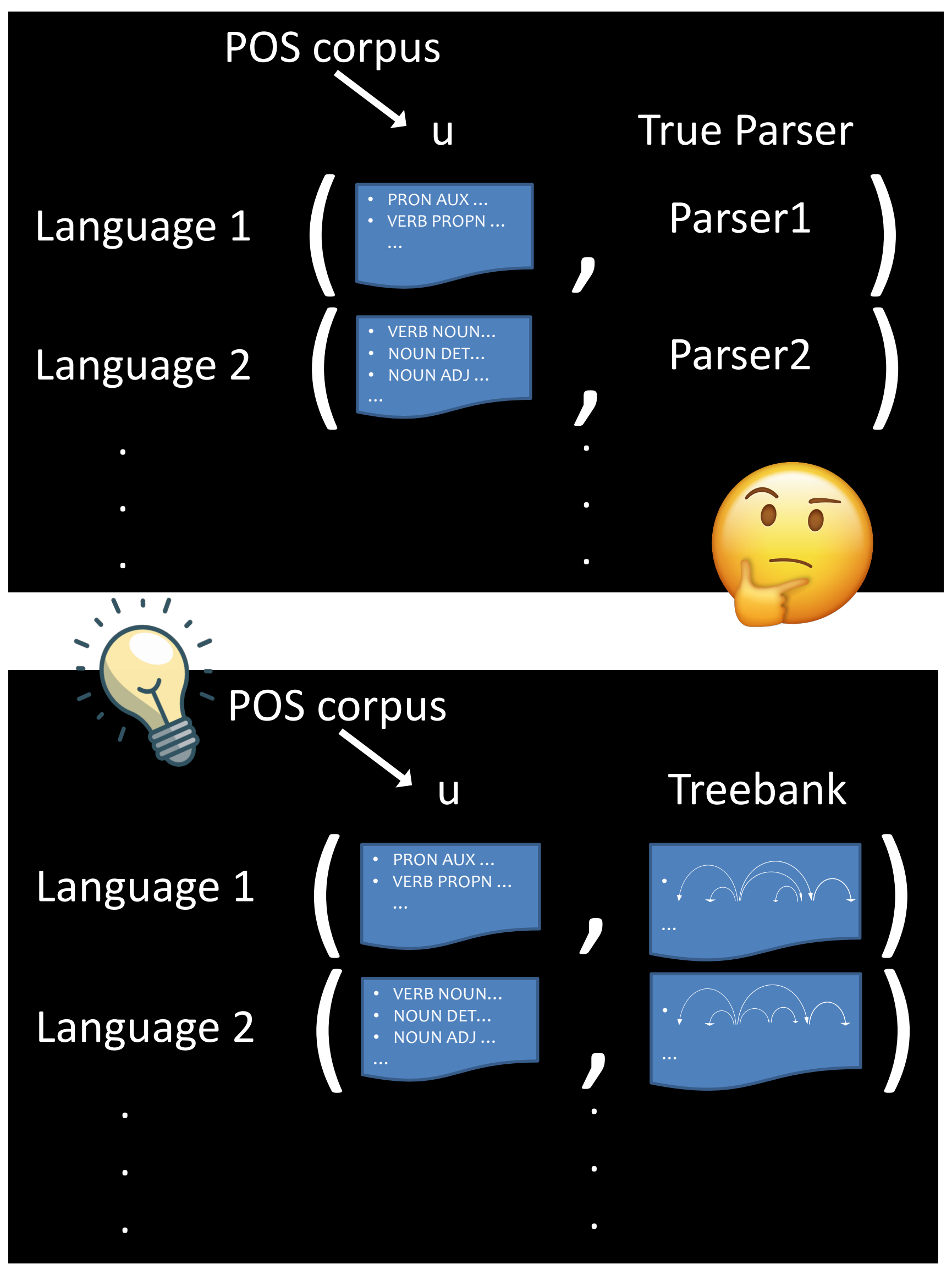
You find a POS-tagged corpus of text in an *unknown* language.

Can you parse this?

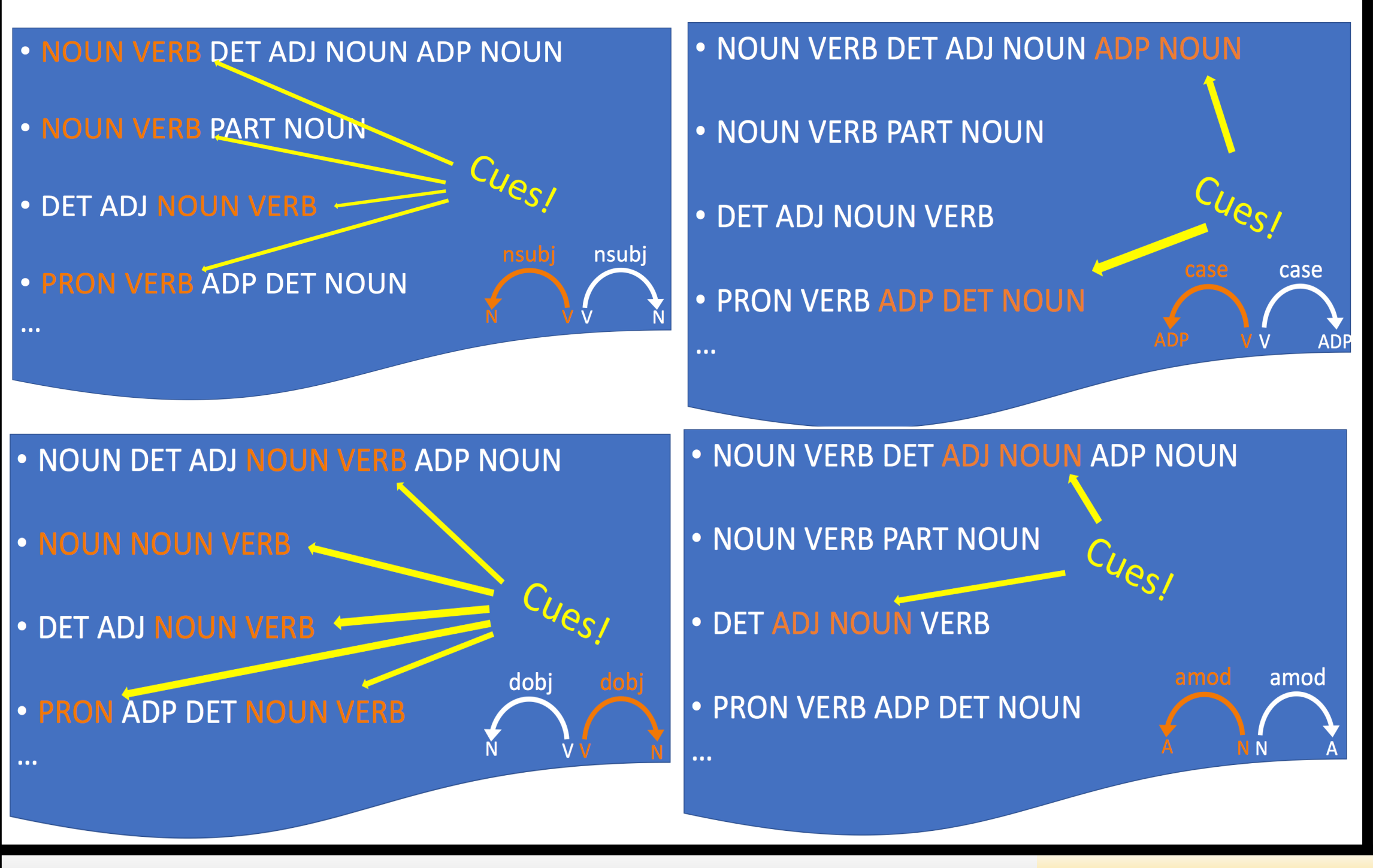
VERB DET NOUN ADJ DET NOUN

- 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.

Each language is an example!

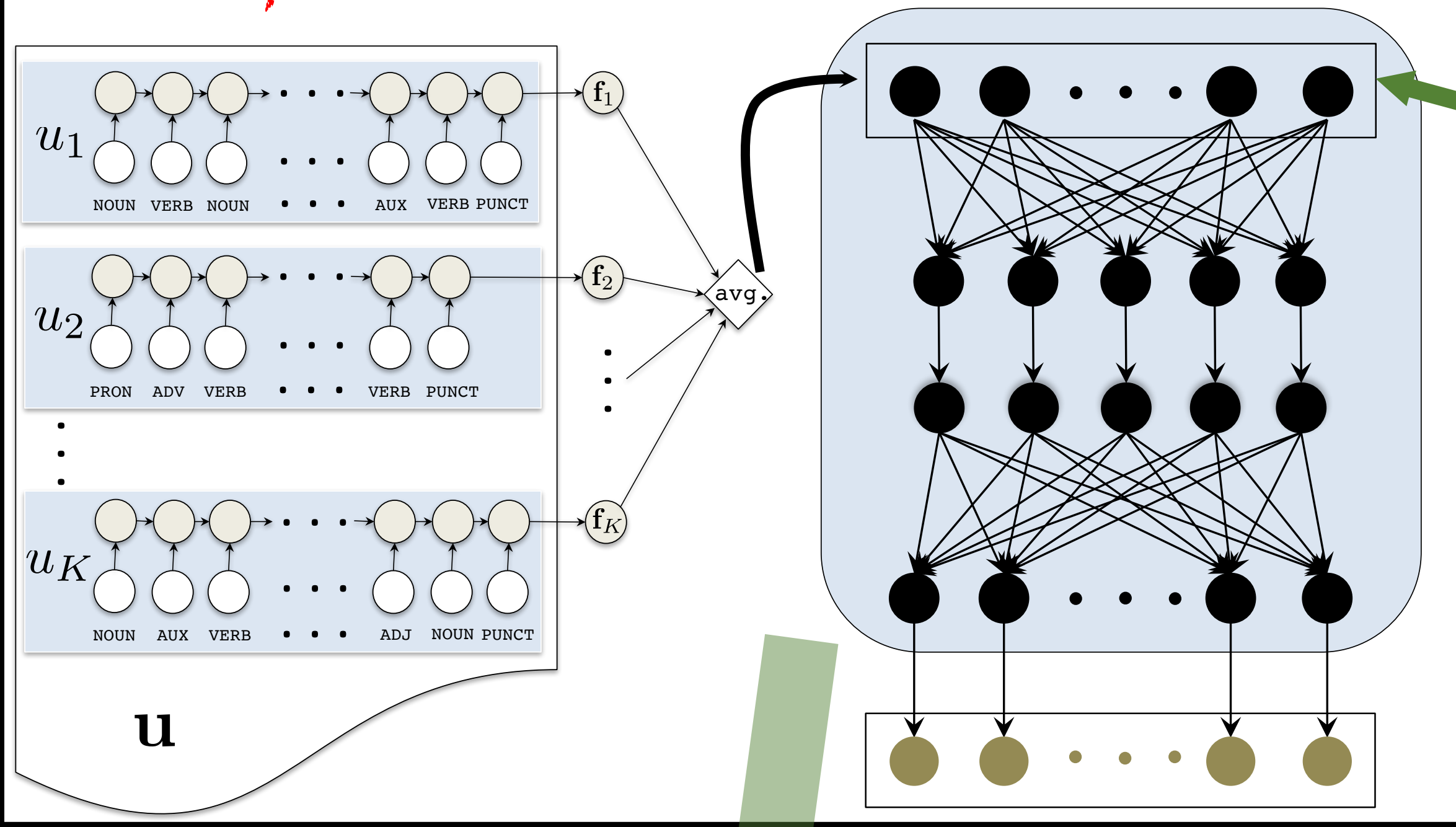

POS corpus
u
True Parser
Language 1
PRON AUX ... VERB PROPN ...
Parser1
Language 2
VERB NOUN... NOUN DET... NOUN ADJ ...
Parser2
POS corpus
u
Treebank
Language 1
PRON AUX ... VERB PROPN ...
Language 2
VERB NOUN... NOUN DET... NOUN ADJ...

Surface Cues to Structure



The Typology Component

Neural features


u_1
NOUN VERB NOUN ... AUX VERB PUNCT
f_1
u_2
PRON ADV VERB ... VERB PUNCT
f_2
u_K
NOUN AUX VERB ... ADJ NOUN PUNCT
f_K
avg.
u

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?

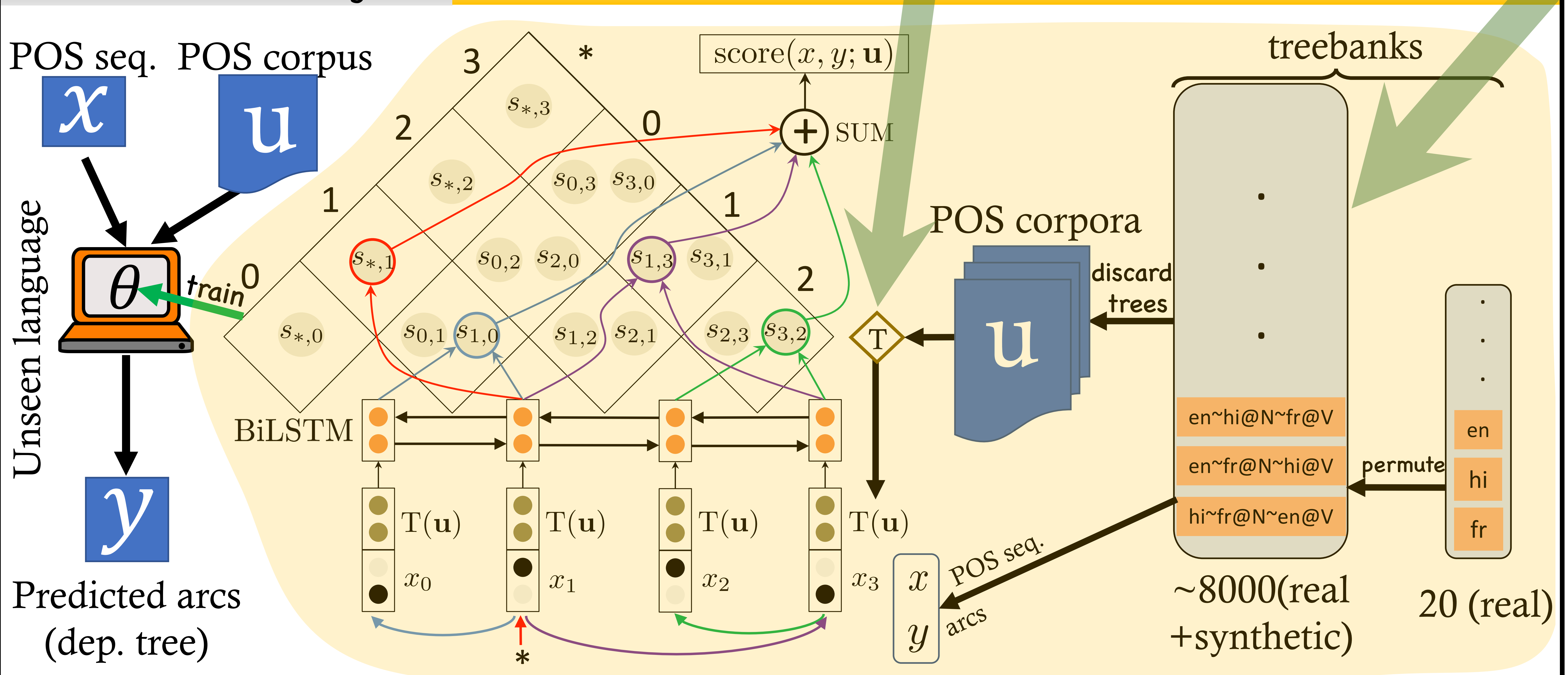
...

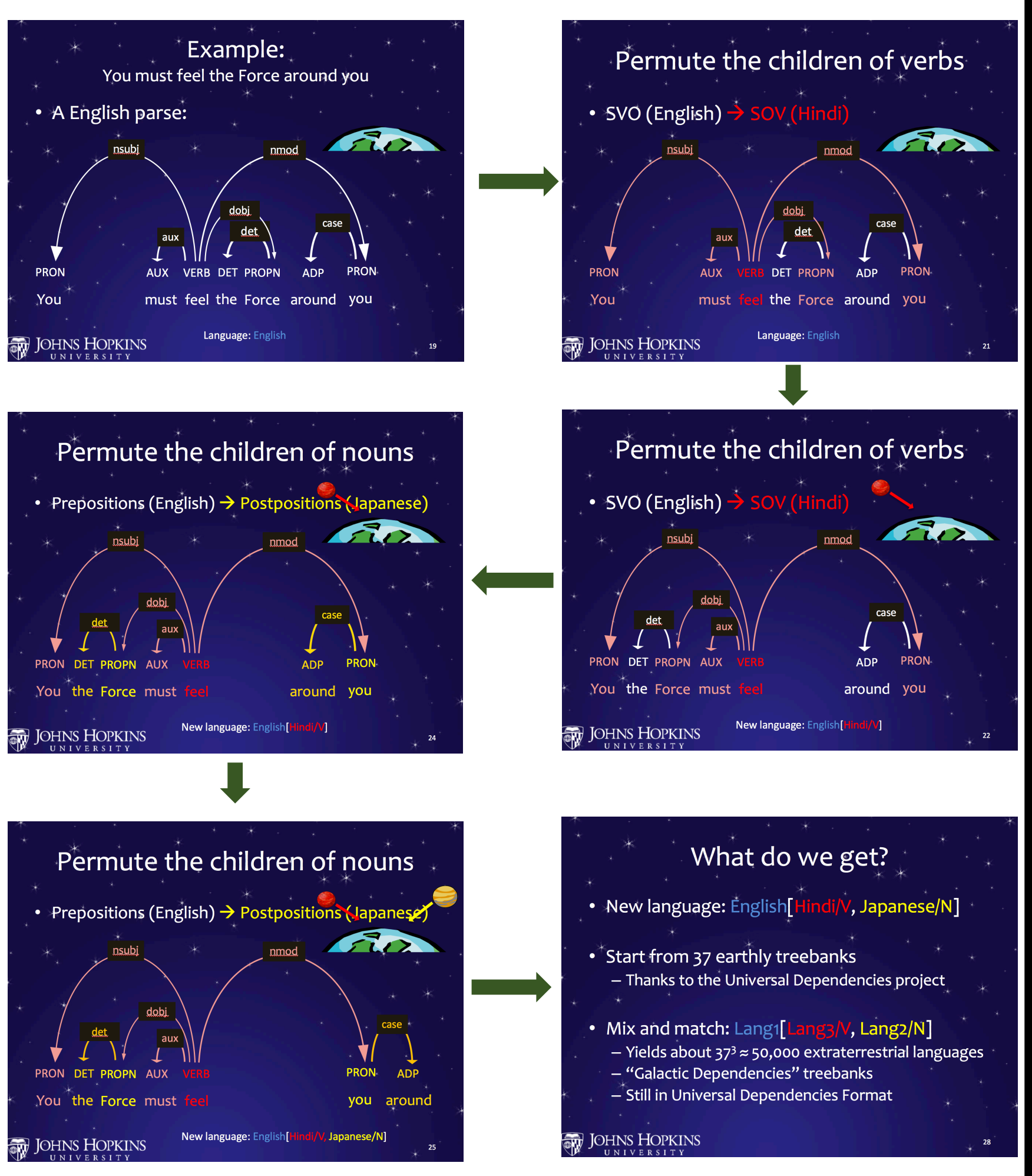
Galactic Treebanks (Wang & Eisner 2016)

- 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

What are we building?

How do we train?




Example:
You must feel the Force around you

- A English parse:

nsubj nmod aux dobj det case
PRON AUX VERB DET PROPN ADP PRON
You must feel the Force around you

Language: English

Permute the children of verbs

- SVO (English) → SOV (Hindi)

PRON AUX VERB DET PROPN ADP PRON
You must feel the Force around you

Language: English

Permute the children of verbs

- SVO (English) → SOV (Hindi)

PRON DET PROPN AUX VERB ADP PRON
You the Force must feel around you

New language: English[Hindi/V]

Permute the children of nouns

- Prepositions (English) → Postpositions (Japanese)

PRON DET PROPN AUX VERB ADP PRON
You the Force must feel around you

New language: English[Hindi/V]

Permute the children of nouns

- Prepositions (English) → Postpositions (Japanese)

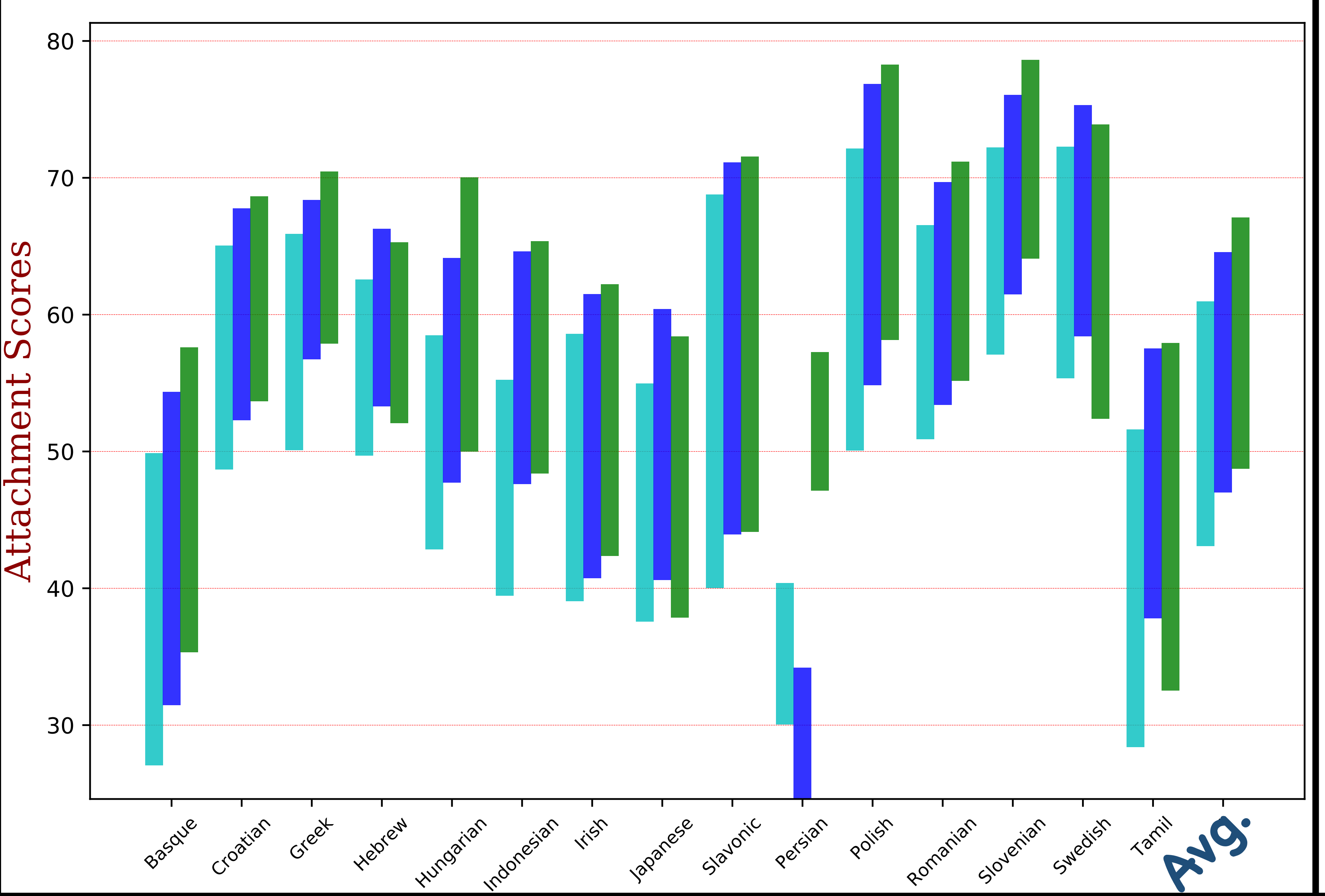
PRON DET PROPN AUX VERB PRON ADP
You the Force must feel you around

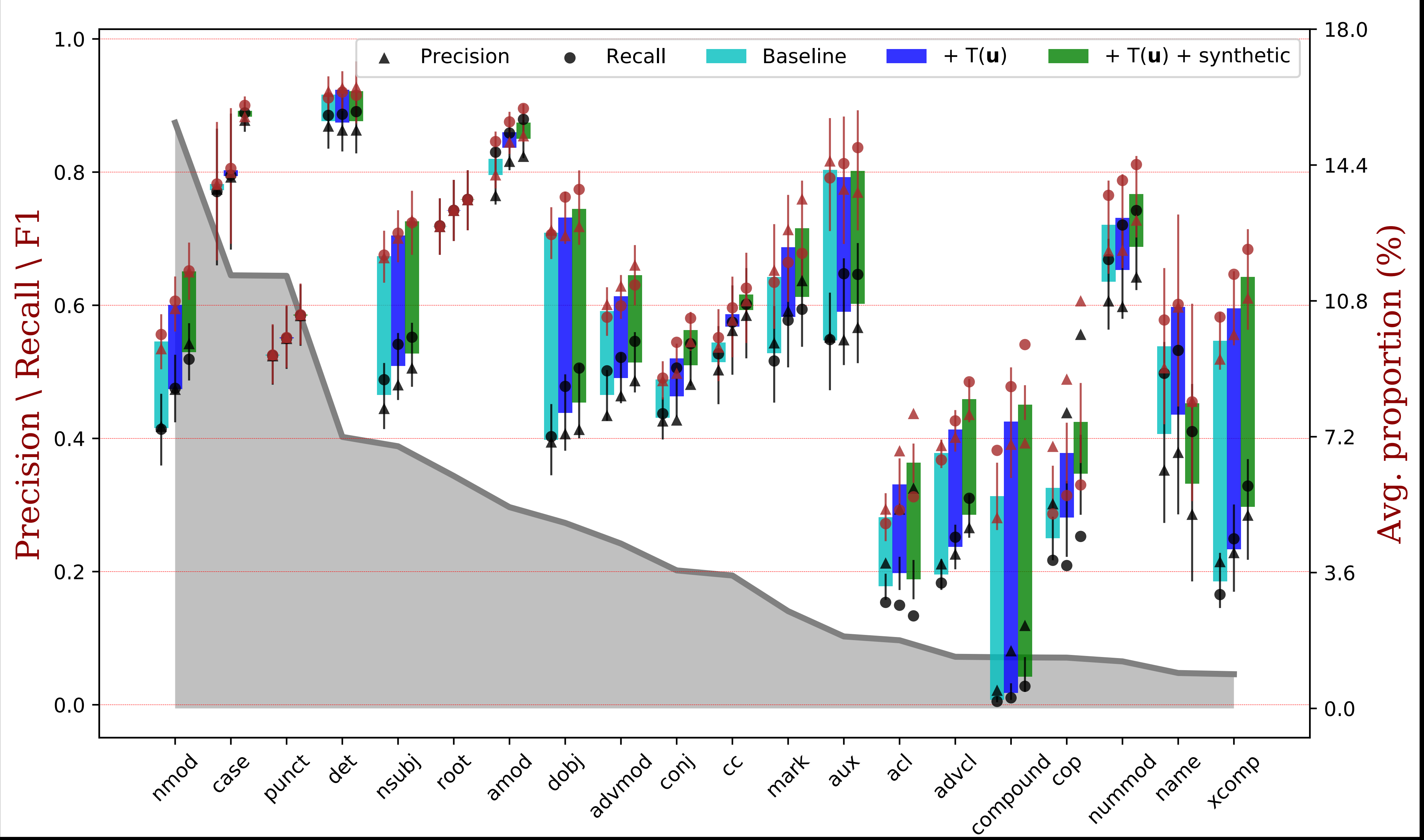
New language: English[Hindi/V, Japanese/N]

What do we get?

- New language: English[Hindi/V, Japanese/N]
- Start from 37 earthly treebanks
  - Thanks to the Universal Dependencies project
- Mix and match: Lang1[Lang3/V, Lang2/N]
  - Yields about $37^3 \approx 50{,}000$ extraterrestrial languages
  - "Galactic Dependencies" treebanks
  - Still in Universal Dependencies Format

Results (each bar stretches from labeled to unlabeled score)


Attachment Scores
Basque
Croatian
Greek
Hebrew
Hungarian
Indonesian
Irish
Japanese
Slavonic
Persian
Polish
Romanian
Slovenian
Swedish
Tamil
Avg.


Precision \ Recall \ F1
Avg. proportion (%)
▲ Precision
● Recall
Baseline
+ T(u)
+ T(u) + synthetic
nmod
case
punct
det
nsubj
root
amod
dobj
advmod
conj
cc
mark
aux
acl
advcl
compound
cop
nummod
name
xcomp