# Predicting Fine-Grained Syntactic Typology from Surface Features

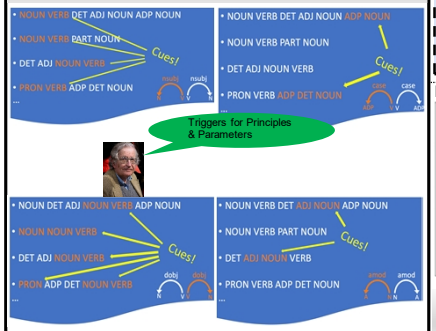
Abstract of Wang and Eisner (2017), previously published in TACL journal & presented at ACL 2017

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

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

We are motivated by the longstanding challenge of determining the structure of a language from its superficial features. Principles & Parameters theory (Chomsky, 1981) hypothesized that human babies are born with an evolutionarily tuned system that is specifically adapted to natural language, which can predict typological properties ("parameters") by spotting telltale configurations in purely linguistic input (Gibson and Wexler, 1994). Here we investigate whether such configurations even exist, by asking an artificial system to find them.

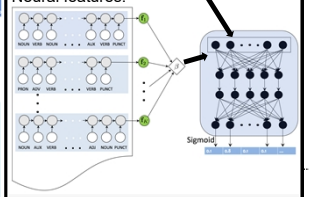
## Surface Cues to Structure




## Hand-Engineered features:

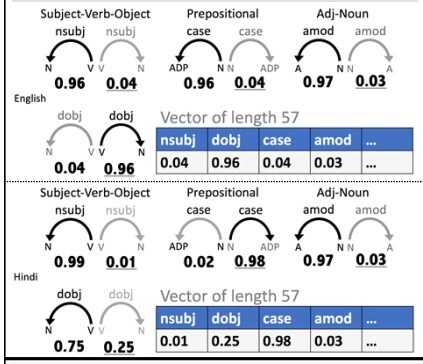
1. How often do NOUNs tend to appear shortly before or after VERBs?
2. How often do ADJs tend to appear shortly before or after NOUNs?
3. How often do ADPs tend to appear shortly before or after NOUNs?
...

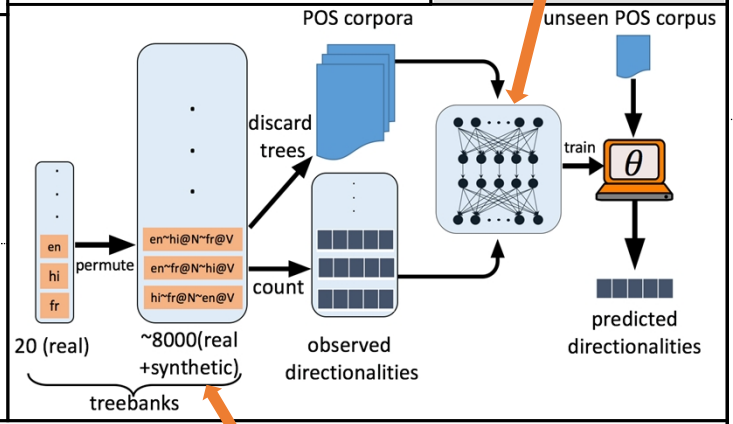
## Neural features:




## Architecture

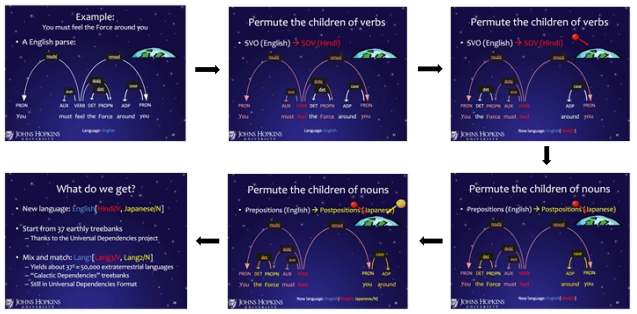
## Fine-Grained Syntactic Typology



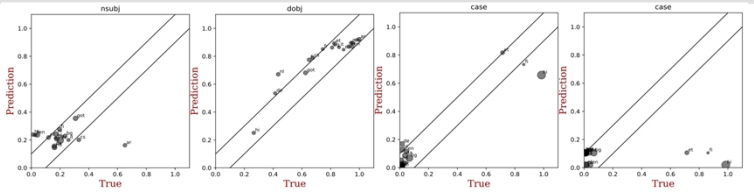




## The Galactic Dependencies Treebanks

- 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

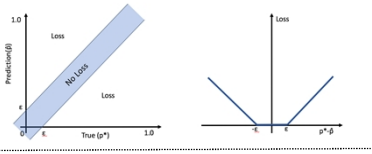


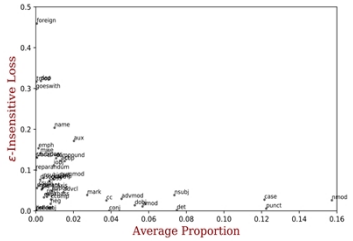

## Results



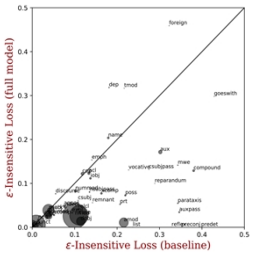

Scatterplots by language of predicted (y-axis) vs. true (x-axis) directionality. The first 3 plots show predictions by our full system on some of the relations. The 4th shows how performance degrades without the use of synthetic data to illustrate the surface word order of postpositional languages.

## ε-insensitive loss




Cross-validation loss broken down by relation. We plot each relation r with x coordinate = the proportion of r in the average training corpus, and with y coordinate = the weighted average ε-insensitive loss.




The y coordinate is the average loss of our model. The x coordinate is the average loss of a simple baseline model that ignores the input corpus.

| | MS13 | N10 | EC | ∅ | UD | +GD |
|---|---|---|---|---|---|---|
| loss | 0.156 | 0.134 | 0.110 | 0.093 | 0.090 | **0.044** |

Cross-validation average expected loss of the two grammar induction methods, MS13 (Mareček and Straka, 2013) and N10 (Naseem et al., 2010), compared to the "expected count" (EC) heuristic and our approach. In these experiments, the dependency relation types are ordered POS pairs. N10 harnesses prior linguistic knowledge, but its improvement upon MS13 is not statistically significant. Both grammar induction systems are *significantly* worse than the rest of the systems, including even our 2 baseline systems, namely EC and ∅ (the no-feature baseline system).

| Family | Sub-Family | Language (Treebank ID) | Split | UD | +GD |
|---|---|---|---|---|---|
| Indo-European | Germanic | Danish (da) | Train1 | 0.024 | 0.017 |
| | | Norwegian (no) | Train1 | 0.008 | 0.011 |
| | | German (de) | Train3 | 0.046 | 0.027 |
| | | Gothic (got) | Train3 | 0.008 | 0.030 |
| | | Dutch (nl) | Train5 | 0.069 | 0.064 |
| | | English (en) | Train5 | 0.025 | 0.036 |
| | | **Swedish (sv)** | **Test** | 0.012 | 0.007 |
| | Slavic | Czech (cs) | Train2 | 0.025 | 0.014 |
| | | Bulgarian (bg) | Train4 | 0.037 | 0.015 |
| | | **Croatian (hr)** | **Test** | 0.062 | 0.012 |
| | | **Old_Church_Slavonic (cu)** | **Test** | 0.024 | 0.029 |
| | | **Polish (pl)** | **Test** | 0.056 | 0.022 |
| | | **Slovenian (sl)** | **Test** | 0.015 | 0.031 |
| | Romance | Portuguese (pt) | Train2 | 0.038 | 0.004 |
| | | Italian (it) | Train3 | 0.011 | 0.010 |
| | | French (fr) | Train4 | 0.024 | 0.020 |
| | | Spanish (es) | Train5 | 0.012 | 0.008 |
| | | **Romanian (ro)** | **Test** | 0.029 | 0.009 |
| | Greek | **Greek (el)** | **Test** | 0.056 | 0.010 |
| | Celtic | **Irish (ga)** | **Test** | 0.181 | 0.159 |
| | Indic | Hindi (hi) | Train5 | 0.363 | 0.173 |
| | Iranian | **Persian (fa)** | **Test** | 0.220 | 0.121 |
| Uralic | Finnic | Estonian (et) | Train2 | 0.055 | 0.015 |
| | | Finnish (fi) | Train4 | 0.069 | 0.070 |
| | Ugric | **Hungarian (hu)** | **Test** | 0.119 | 0.101 |
| Afro-Asiatic | Semitic | Arabic (ar) | Train1 | 0.116 | 0.056 |
| | | **Hebrew (he)** | **Test** | 0.079 | 0.034 |
| Austronesian | - | **Indonesian (id)** | **Test** | 0.099 | 0.073 |
| Basque | - | **Basque (eu)** | **Test** | 0.250 | 0.077 |
| Dravidian | Southern | **Tamil (ta)** | **Test** | 0.238 | 0.052 |
| Japanese | - | **Japanese (ja_ktc)** | **Test** | 0.247 | 0.080 |
| Average | | | **Test** | 0.112 | 0.054* |
| | | | Train+Test | 0.084 | 0.045* |

Our final comparison on the 15 test languages (boldfaced). We ask whether the average expected loss on these 15 real target languages is reduced by augmenting the training pool of 20 UD treebanks with +20*21*21 GD languages. For completeness, we extend the table with the cross-validation results on the training pool. The "Avg." lines report the average of 15 test or 31 training+testing languages. We mark both "+GD" averages with "*" as they are significantly better than their "UD" counterparts (paired permutation test by language, $p < 0.05$).