Phylogenetic Multi-Lingual Dependency Parsing
(NAACL 2019)

Zhao Li
2019.6.12

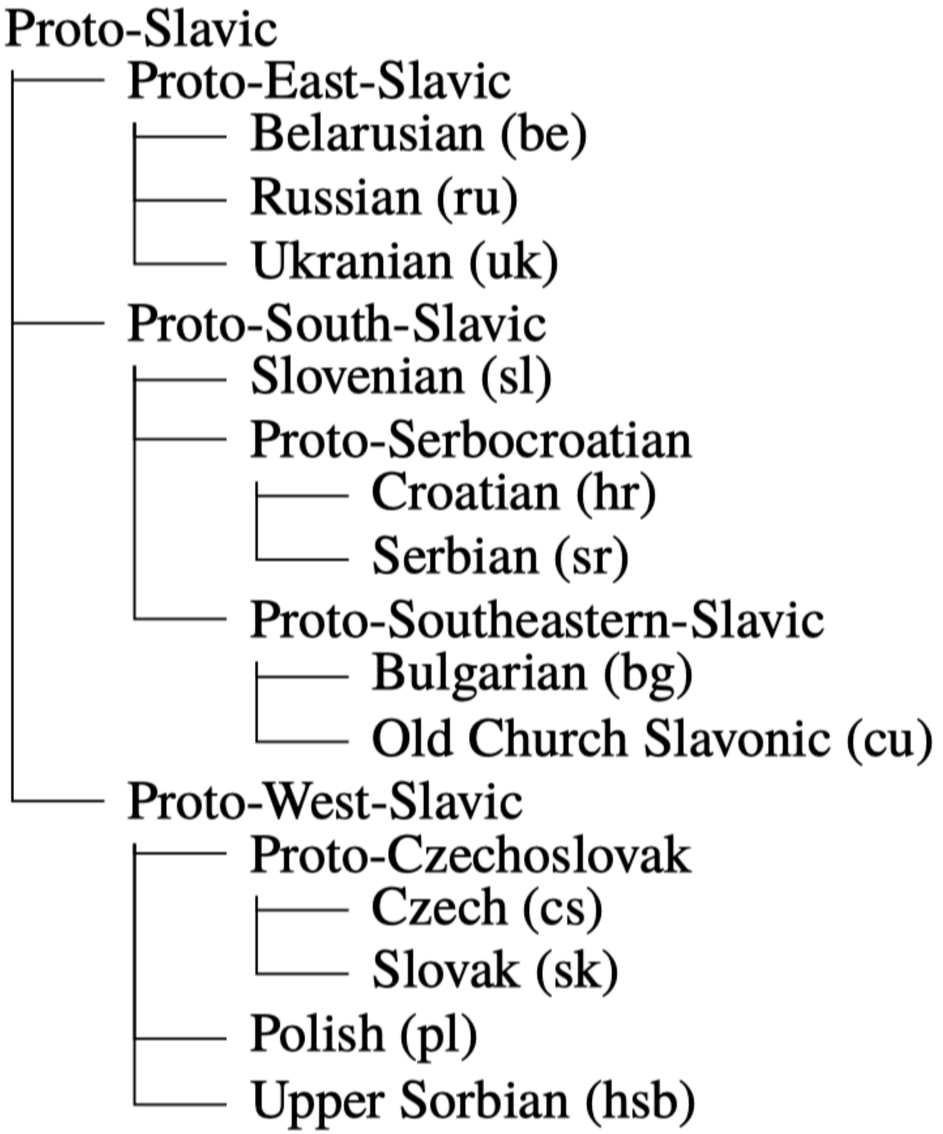

Figure 1: A possible phylogenetic tree for languages in the Slavic family.

Motivation

- Languages evolve and diverge over time. Their evolutionary history is often depicted in the shape of a phylogenetic tree.

- Assuming parsing models are representations of their languages grammars, their evolution should follow a structure similar to that of the phylogenetic tree.

Contribution

- Make use of the phylogenetic tree to guide the learning of multi-lingual dependency parsers leveraging languages structural similarities.
- Experiments show that phylogenetic training is beneficial to low resourced languages and to well furnished languages families.
- Phylogenetic training is able to perform zero-shot parsing of previously unseen languages.

Phylogenetic Hypothesis

- Assume that the grammar of the last common ancestor is a good approximation of those languages grammars.
- Issues with this assumption: **1.** A language grammar can be very different from its ancestor one from two millennia earlier. **2.** A lot of languages have only started to be recorded very recently thus lacking historical data all together.
- Solution: Use all the data from descendent languages to represent an ancestor language.

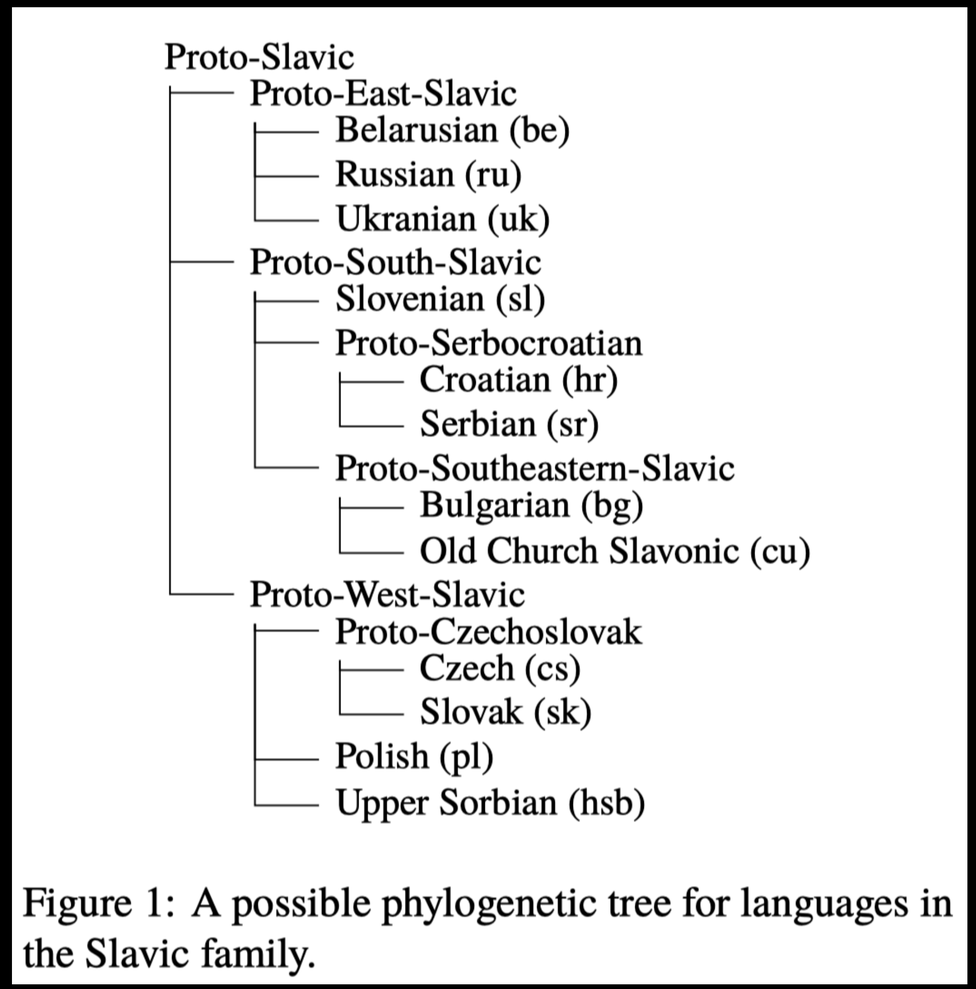

Figure 1: A possible phylogenetic tree for languages in the Slavic family.

Neural Model

$$\mathbf{w}_i = \mathbf{pos}_i \oplus \mathbf{morph}_i \oplus \mathbf{char}_i.$$

$$\mathbf{morph}_i = \sum_{m \in morph_i} \mathbf{m}_m.$$

Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin

$$\mathbf{c}_i = forward(\mathbf{w}_1, ..., \mathbf{w}_i) \oplus backward(\mathbf{w}_i, ..., \mathbf{w}_l).$$

$$s_{ij} = \max_l s_{ijl} = \max_l (\mathbf{L}_2 \cdot [\mathbf{L}_1 \cdot (\hat{\mathbf{c}}_i \oplus \mathbf{c}_j)]^+)_l.$$

$$loss(x) = \sum_{w_i} \Big[\sum_{\substack{j' \neq j \\ j' \neq i}} \max(0, s_{ij'} - s_{ij} + 1)^2 + \sum_{l' \neq l} \max(0, s_{ijl'} - s_{ijl} + 1)^2 \Big]$$

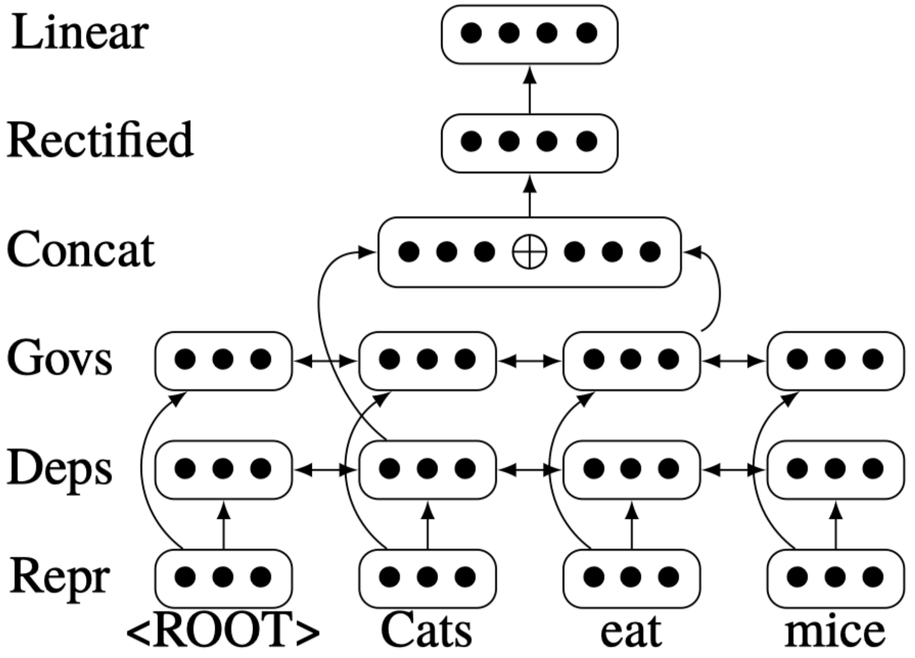

Figure 3: Neural network architecture for edge scoring. The contextualised representation of the governor (eat) and the dependent (Cats) are concatenated and passed through a rectified linear layer and a final plain linear layer to get a vector of label scores.

$$\mathcal{L} = \{l_1, l_2, ..., l_{n_l}\}$$

$$\mathcal{P} = \{p_1, p_2, ..., p_{n_p}\}$$

Let $\mathcal{T}$ be a tree over $\mathcal{L}^* = \mathcal{L} \cup \mathcal{P}$

notation $p > l$

language $l \in \mathcal{L}$, a set of n annotated examples $\mathcal{D}_l$.

proto- language $p \in \mathcal{P}$ set $\mathcal{D}_p =$

$\bigcup_{p>l} \mathcal{D}_l$ as the union of its descendent sets.

- The main idea behind phylogenetic training is to initialize a new model with the model of its parent, thus effectively sharing information between languages and letting models diverged and specialize over time.

Data: a train set $\mathcal{D}_l$ and a dev set $\mathcal{D}'_l$ per language, a tree $\mathcal{T}$, two sampling sizes k, k' and a maximum number of reboot r
Result: a model $\boldsymbol{\theta}$ per node in $\mathcal{T}$
begin
  Instantiate empty queue Q
  $Q.push(\mathcal{T}.root)$
  while *Q is not empty* **do**
    $l = Q.pop()$
    if $l = \mathcal{T}.root$ **then**
      initialize $\boldsymbol{\theta}^0_{\mathcal{T}.root}$ randomly
    else
      $\boldsymbol{\theta}^0_l = \boldsymbol{\theta}_{l.parent}$
    $reboot = 0, i = 1, a_0 = 0$
    while $reboot < r$ **do**
      $\boldsymbol{\theta}^i_l = train(\boldsymbol{\theta}^{i-1}_l, \mathcal{D}_l, k)$
      $a_i = test(\boldsymbol{\theta}^i_l, \mathcal{D}'_l, k')$
      if $a_i \leq a_{i-1}$ **then**
        $reboot \mathrel{+}= 1$
      else
        $reboot = 0, i \mathrel{+}= 1$
    $\boldsymbol{\theta}_l = \boldsymbol{\theta}^i_l$
    for *c in l.children* **do**
      $Q.push(c)$

Algorithm 1: Phylogenetic training procedure.

How to sample sentences?

- Sampling sentences uniformly across languages is not a viable option for the size of datasets varies a lot across languages and that they do not correlate with how close a language is to its ancestor.

- In this work, at a given inner node, we decided to sample uniformly at random over branches spanning from this node, then uniformly at random over languages and then uniformly at random over sentences.

- It boils down to flattening the subtree below an inner node to have a maximum depth of 2.

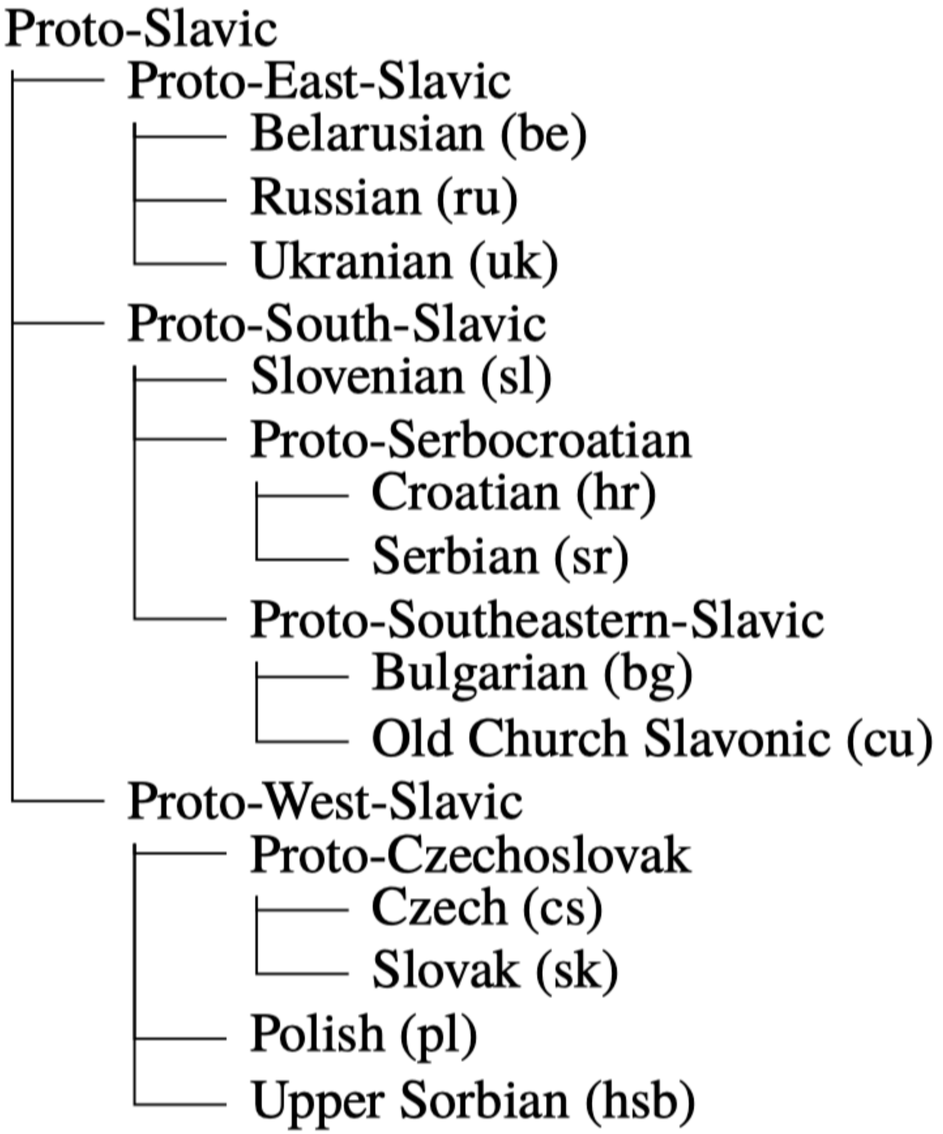

Figure 1: A possible phylogenetic tree for languages in the Slavic family.

Zero-Shot Parsing

- Phylogenetic training procedure provides a model for each inner node of the tree and thus each intermediary grammar.
- If one were to bring a new language with its position in the tree, then we can use the pretrained model of its direct ancestor as an initialization instead of learning a new model from scratch.

Multi-Task Learning

	Phylogenetic		Independent	
	UAS	LAS	UAS	LAS
ar nuyad	74.81	70.32	**75.07**	**71.08**
cop	85.51	79.28	**86.03**	**80.15**
he	**81.89**	75.36	81.59	**75.57**
bxr [19]	**48.72**	**30.68**	37.88	18.09
eu	76.81	69.51	**78.61**	**72.76**
af	85.15	80.94	**85.44**	**81.66**
da	78.50	72.50	**79.16**	**74.13**
de gsd	**80.37**	**73.54**	79.48	72.37
en ewt	79.25	74.34	**79.27**	**74.66**
got	77.83	71.54	**79.91**	**74.33**
nb	**84.62**	**78.78**	83.82	78.09
nl alpino	**77.19**	**68.55**	76.52	68.40
nn nynorsk	82.39	76.44	**82.58**	**77.32**
sv talbanken	80.46	74.62	**81.17**	**75.47**
be	**80.18**	**74.11**	78.09	72.76
bg	86.01	79.16	**86.40**	**79.79**
cs pdt	**79.78**	**71.71**	77.45	69.88
cu	82.98	77.19	**83.31**	**78.32**
hr	**81.70**	**74.73**	81.05	73.95
hsb [23]	**74.24**	**66.01**	58.59	50.37
lt	**61.42**	**50.88**	56.35	46.14
lv	**78.39**	**70.14**	76.69	68.89
pl lfg	**92.88**	**88.53**	91.07	86.49
ru syntagrus	**77.91**	72.72	77.33	**72.85**
sk	**84.91**	**79.17**	79.09	73.20
sl ssj	87.15	83.43	**88.39**	**85.21**
sr	85.85	79.86	**86.17**	**80.47**
uk	**78.16**	**73.50**	74.96	70.91
ca	84.67	78.81	**85.69**	**80.11**
es ancora	85.11	79.52	**85.61**	**80.18**
fr gsd	**84.35**	77.59	84.21	**77.94**
fro	**82.32**	**74.24**	78.91	69.95
gl [600] treegal	**83.80**	**78.06**	83.60	77.63
it isdt	87.03	81.67	**87.10**	**82.27**
la proiel	**66.25**	**58.88**	65.07	57.80
pt bosque	**84.93**	79.37	84.90	**79.83**
ro rrt	79.83	70.46	**79.93**	**70.88**
fa	78.76	72.95	**79.93**	**74.07**
hi hdtb	**89.32**	**82.89**	88.75	82.60
kmr [20]	**69.08**	**59.64**	54.77	45.07
mr	**78.65**	**68.97**	76.60	64.04
ur	84.32	77.02	**84.82**	**78.19**
el	86.44	83.30	**86.88**	**83.96**
grc proiel	**73.82**	**67.88**	71.68	66.05
ga [566]	75.91	67.54	**76.20**	**67.72**
hy [50]	**65.03**	**51.76**	59.27	46.67
id gsd	**81.08**	**74.97**	80.83	74.69
ja gsd	91.22	87.31	**91.40**	**87.37**
ko kaist	73.38	68.35	**74.23**	**69.81**
kk [31]	**70.82**	**55.42**	62.81	44.59
tr imst	**59.64**	**50.66**	59.00	50.54
ug	**66.33**	**48.20**	63.66	46.07
et	**75.32**	**68.13**	73.91	66.96
fi ftb	**78.05**	**72.20**	74.66	68.22
hu	79.51	72.88	**80.15**	**74.31**
sme [2257]	**80.13**	**76.40**	78.34	74.25
ta	75.05	66.94	**76.19**	**67.93**
te	**88.88**	**74.24**	87.01	72.05
vi	65.59	61.15	**66.02**	**61.74**
zh	**80.36**	**74.79**	80.14	74.52
Avg	80.05	73.35	79.47	73.02
Avg No Dev	70.97	60.69	63.93	53.05

Table 1: Parsing results for languages with a training set for phylogenetic models and independent models. The training set size of languages without a developpement set are reported in brackets.

Zero-Shot Parsing

Lang	Model	UAS	LAS
am	Semitic	57.27	26.25
br	Celtic*	61.36	43.89
fo	North-Germanic	52.40	46.52
sa	Indic	56.18	40.46
kpv lattice	Finno-Permiac*	65.16	52.11
pcm	World	60.43	43.80
th	World	29.14	17.61
tl	Austronesian*	70.89	50.38
wbp	World	87.67	65.66
yo	World	56.16	37.51
yue	Sino-Tibetan*	41.68	25.02
Avg		58.04	40.83

Table 2: Accuracy of languages without a training set.