Typological Feature Prediction and Blinding for Cross-Lingual NLP

Johannes Bjerva
Department of Computer Science
Aalborg University Copenhagen

AALBORG UNIVERSITY

Outline

- Introducing “The Problem”
  - Typological Feature Prediction
- Is Typology “The Solution”?
  - NLP Models **Encode** Typology
  - NLP Models **Need** Typology
  - ...but does adding more Typology help?

The Big Problem

Modern society relies on **NLP tools**,

but **most languages** do not have access to such tools,

nor the **resources** or **data** to create them

The Big Problem is Really a Problem

Target

9.c

Significantly increase access to information and communications technology and strive to provide universal and affordable access to the Internet in least developed countries by 2020

Indicators

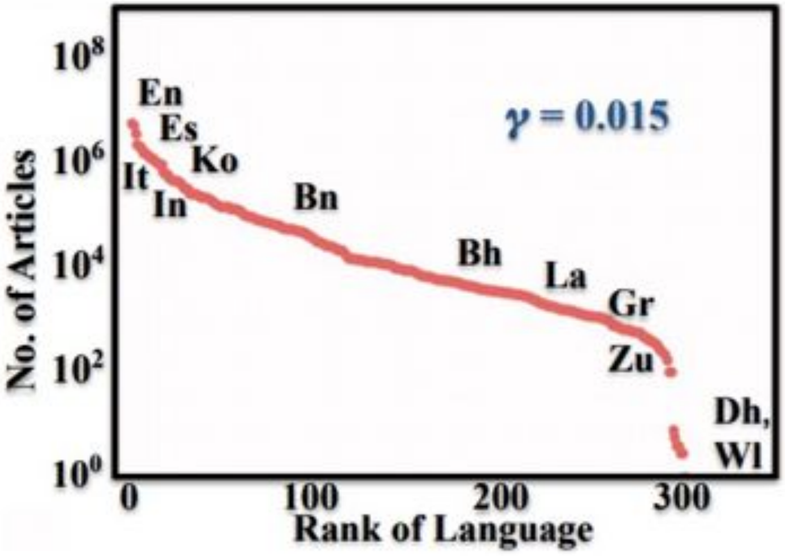
9.c.1

Proportion of population covered by a mobile network, by technology

Cause: Languages are Different

[Joshi et al., ACL2020]

Cause: Multilingual Data is Scarce



(c) Wikipedia

Is Typology the Solution?

red apple

pomme rouge

"Adjective-Noun Order"

Typological Feature Prediction

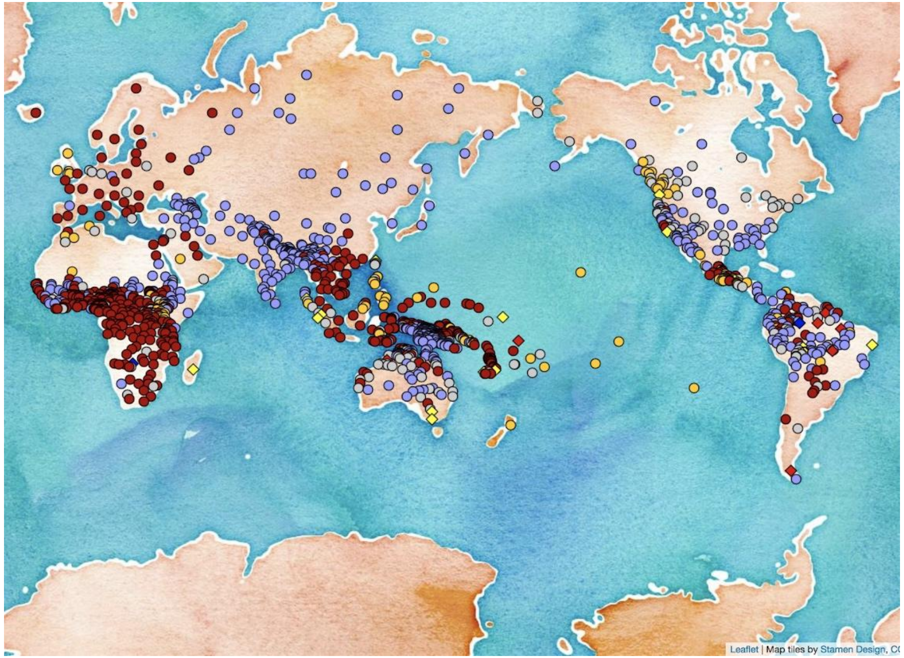
Feature 81A – Order of Subject, Object and Verb

World Atlas of Language Structure (WALS)

- 2,500 languages
- 192 features

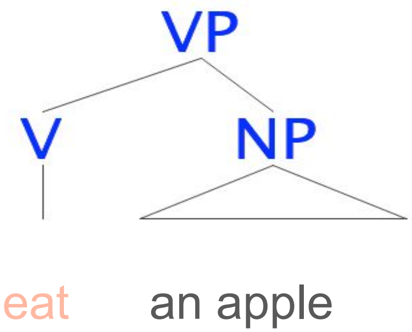
Values

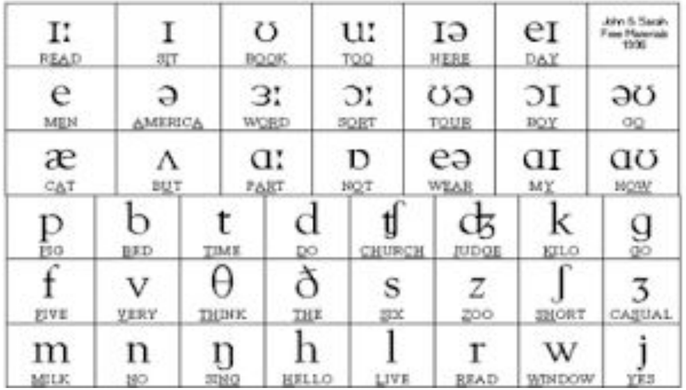
	SOV	565
	SVO	488
	VSO	95
	VOS	25
	OVS	11
	OSV	4
	No dominant order	189

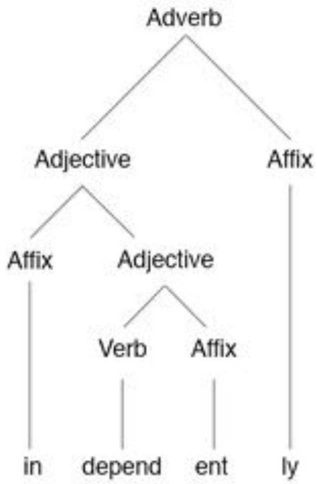


Typological Features

- Syntax
- Morphology
- Phonology
- …



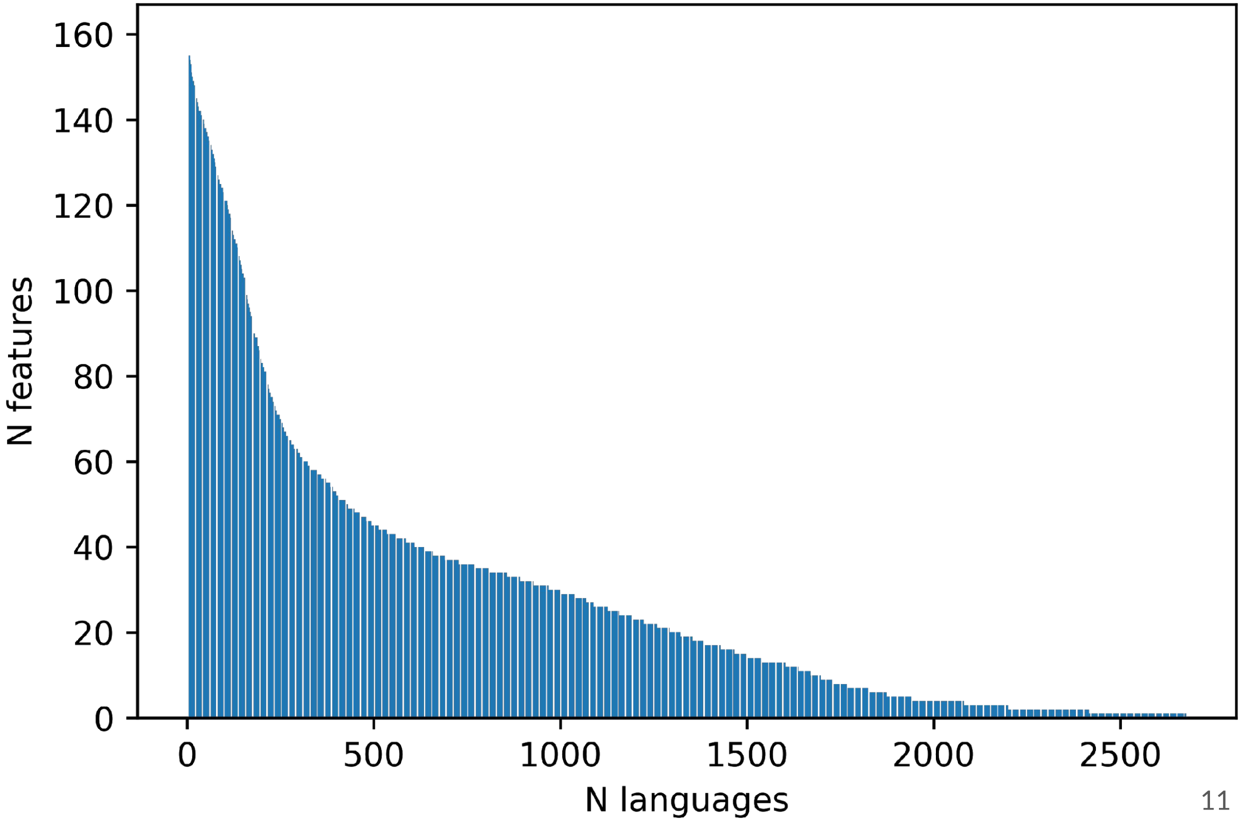




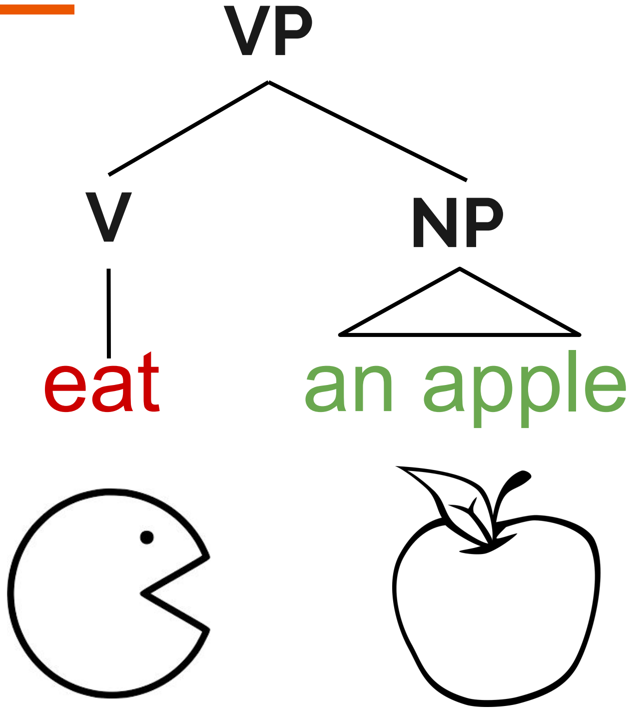
Motivation: Typological Databases are Incomplete

- **Sparse**:
  Most languages are covered by only a handful of features

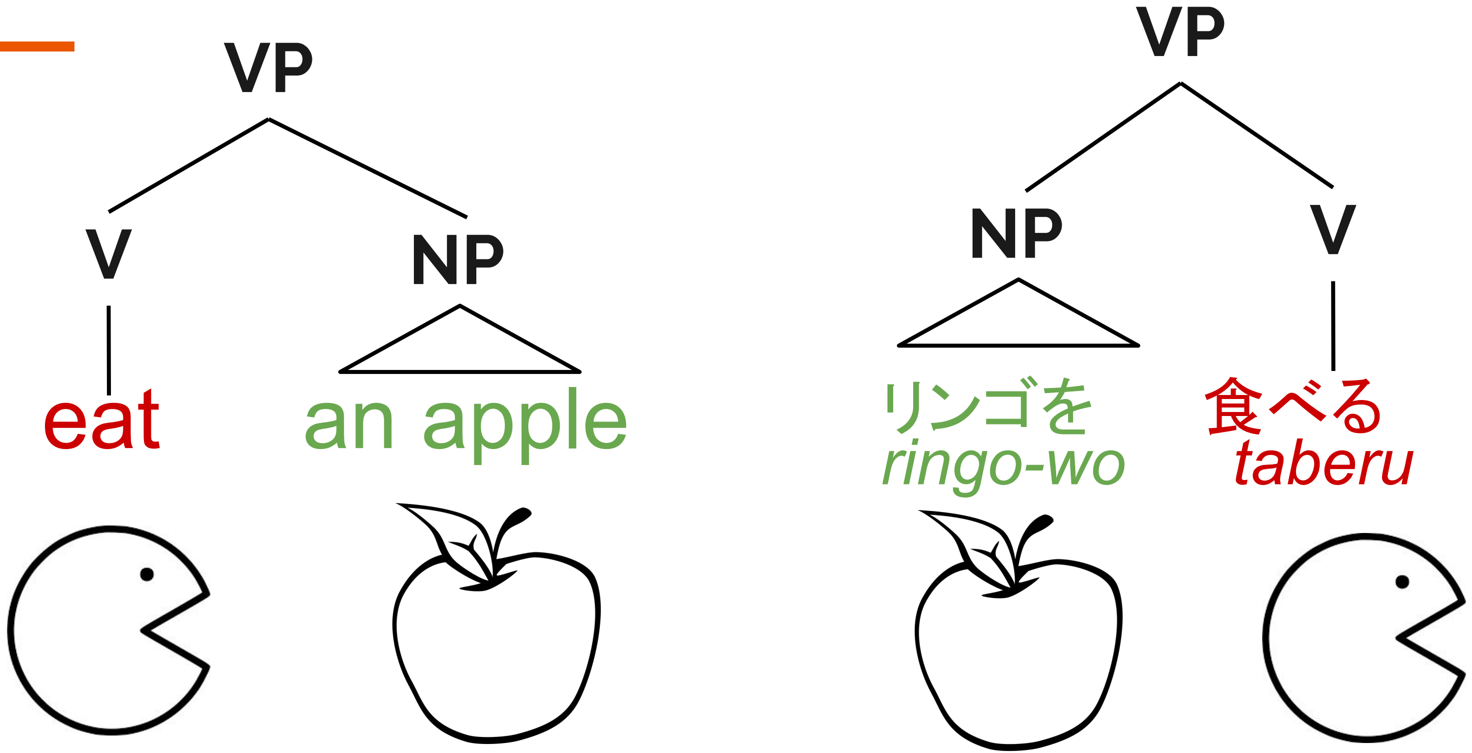
- **Skewed**:
  A few features have much wider coverage than others



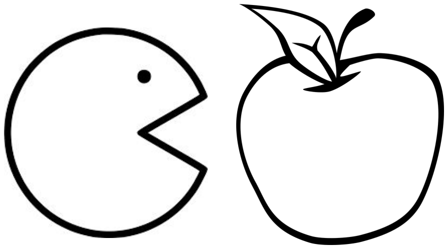
Languages are different

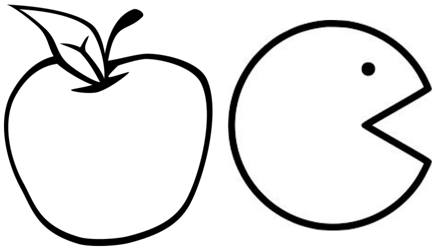


Languages are different



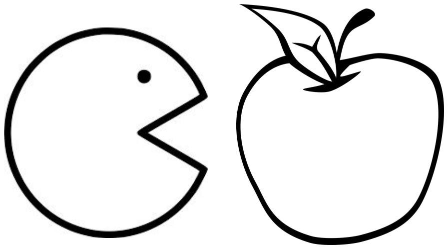
Differences are often correlated

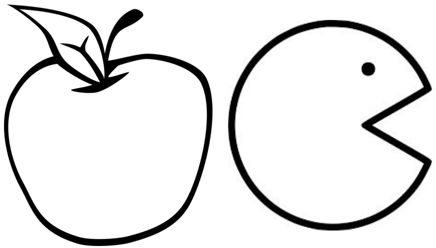


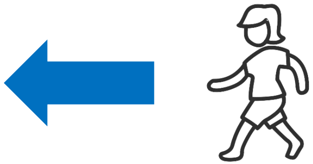


go to SIGTYP

Differences are often correlated







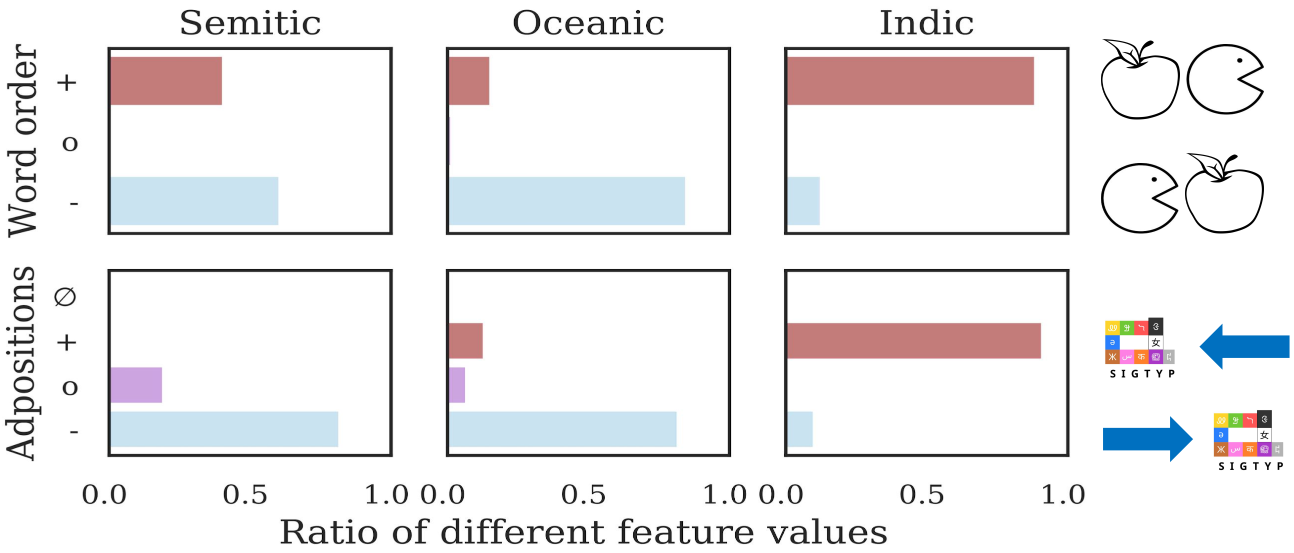
go to SIGTYP

SIGTYP に 行く

SIGTYP -ni iku

[Bjerva, Kementchedjhieva, Cotterell, Augenstein (NAACL 2019)]

Typological Feature Correlations



[Bjerva, Salesky, Mielke, Chaudhary, Celano, Ponti, Vylomova, Cotterell, Augenstein (SIGTYP 2020)]

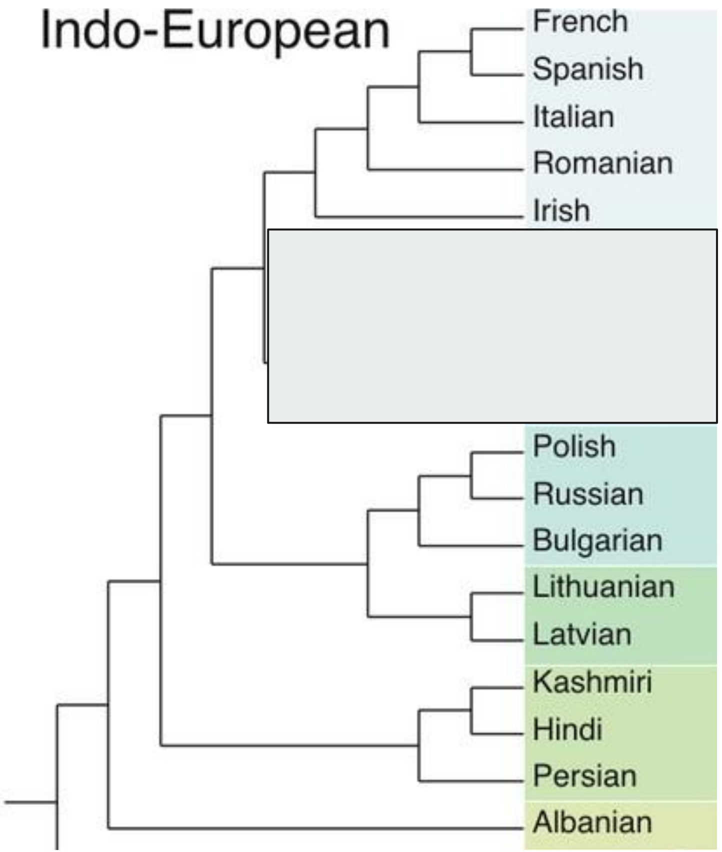
Typological Feature Prediction

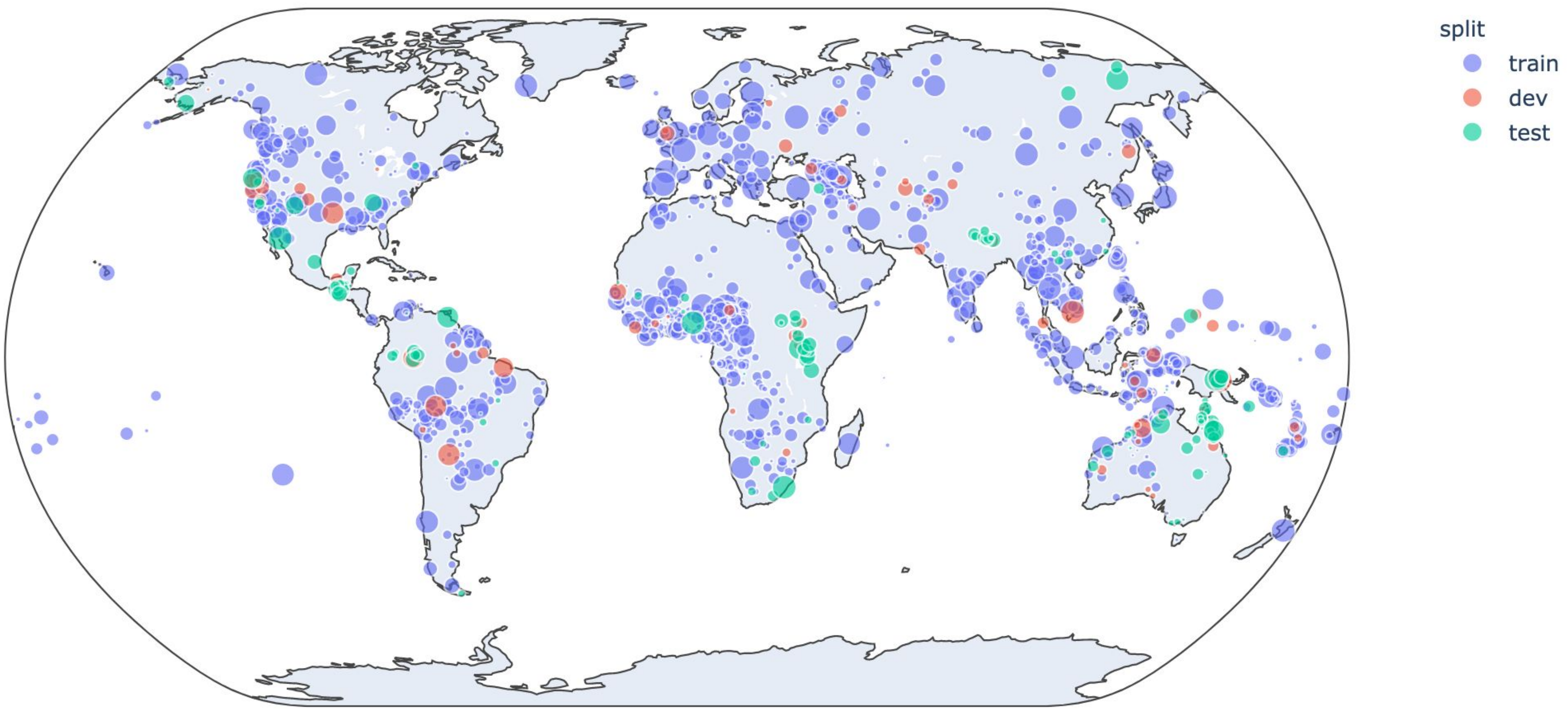
	Lang code	Name	Lat	Long	Genus	Family	Count Code	Features
Input	mhi	Marathi	19.0	76.0	Indic	Indo-European	IN	order_of_subject,_object,_and_verb=? | number_of_genders=?
	jpn	Japanese	37.0	140.0	Japanese	Japanese	JP	case_syncretism=? | order_of_adjective_and_noun=?
Output	mhi	Marathi	19.0	76.0	Indic	Indo-European	IN	order_of_subject,_object,_and_verb=SOV | number_of_genders=three
	jpn	Japanese	37.0	140.0	Japanese	Japanese	JP	case_syncretism=no_case_marking | order_of_adjective_and_noun=demonstrative-Noun

Table 1: Data format for two test instances of the SIGTYP 2020 shared task dataset

Typological Feature Prediction

- Not difficult with knowledge of related languages
- Controlling for Phylogenetic Relationships
  - Evaluation on held-out language families
- Controlling for Geographic Influence
  - Training languages > 1000 km
    from evaluation languages

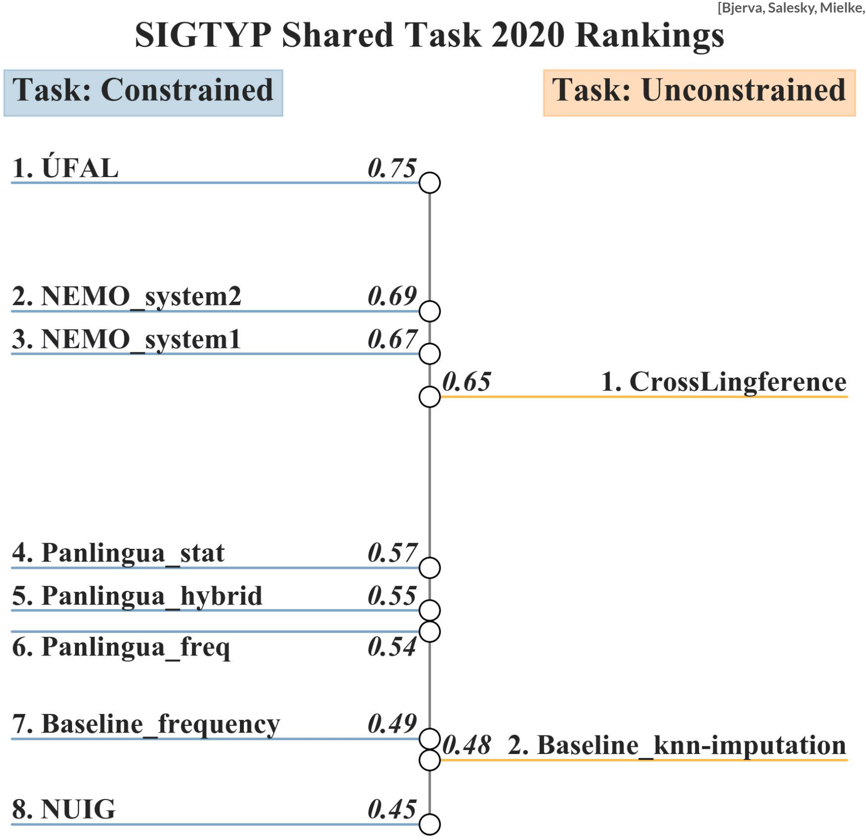




split
train
dev
test

Results

[Bjerva, Salesky, Mielke, Chaudhary, Celano, Ponti, Vylomova, Cotterell, Augenstein (SIGTYP 2020)]



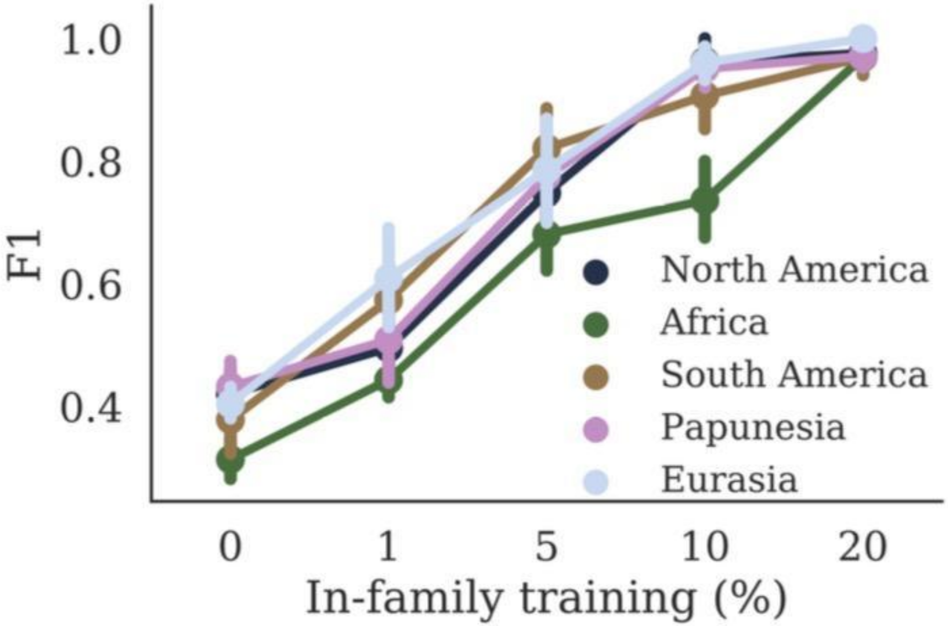
[Bjerva, Salesky, Mielke, Chaudhary, Celano, Ponti, Vylomova, Cotterell, Augenstein (SIGTYP 2020)]

Submission	**Tucanoan** (8)	**Madang** (9)	**Mahakiranti** (13)	**Nilotic** (15)	**Mayan** (17)	**N. Pama-Nyungan** (24)	**Other genera** (63)
ÚFAL	**0.73**	**0.78**	**0.74**	0.71	**0.80**	**0.76**	**0.76**
NEMO_system2	0.71	0.72	0.72	**0.76**	0.76	0.67	0.69
NEMO_system1	0.70	0.72	0.68	0.75	0.71	0.68	0.67
Panlingua_stat	0.70	0.64	0.55	0.55	0.33	0.62	0.58
Panlingua_hybrid	0.65	0.64	0.57	0.51	0.34	0.61	0.53
Panlingua_freq	0.59	0.64	0.53	0.55	0.31	0.59	0.55
Baseline_frequency	0.51	0.53	0.37	0.49	0.41	0.58	0.53
NUIG	0.51	0.56	0.35	0.45	0.32	0.45	0.48
CrossLingference	0.71	0.73	0.67	0.68	0.57	0.60	0.65
Baseline_knn-imputation	0.48	0.57	0.46	0.48	0.32	0.52	0.51

Table 2: Macro-averaged results across each unobserved genus, as compared to genera with languages observed in training with randomly sampled splits, shown with number of languages in each genus.

[Bjerva, Kementchedjhieva, Cotterell, Augenstein (NAACL 2019)]

Sufficient In-family Training



Typological Feature Prediction: Conclusion

- “Solved” when we have knowledge of related languages
- Difficult in the SIGTYP 2020 setting
  - Rare features are difficult
    - Best systems: > 65% accuracy
    - Worst systems: ~20% accuracy
- Future work:
  - Explainable modelling
  - More useful to point at real text examples
  - ...certainly easier to convince typologists that way

The Big Problem:
Is Typology the Solution?

Overview

- Is typology the solution?
    - NLP models **encode** typology
    - NLP models **rely on** typology
    - ...so how about **adding** some more typology to our models?

[*SEM 2021]

Is Typology the Solution?

she's eating a red apple

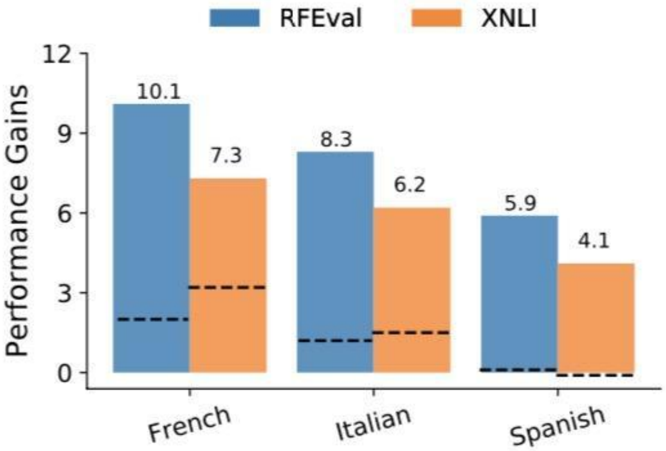
elle mange une pomme rouge

[*SEM 2021]

Is Typology the Solution?

she's eating a red apple

elle mange une rouge pomme

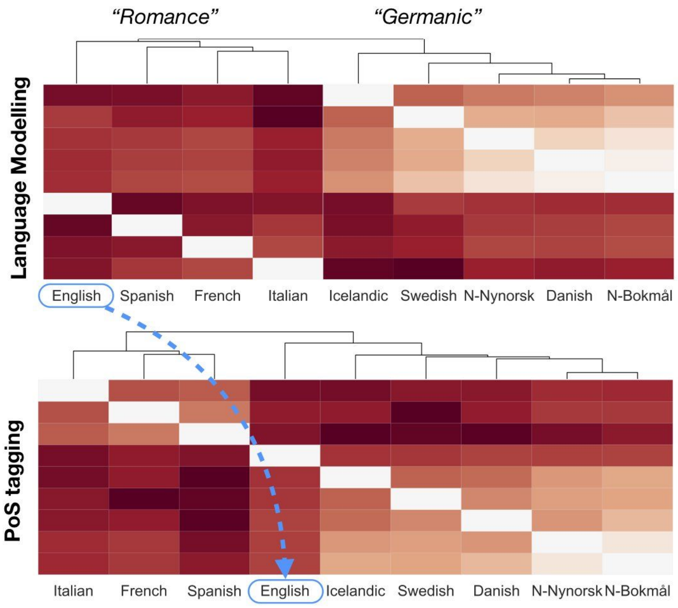


(b) Reversing adjective-noun order

[Bjerva and Augenstein (NAACL 2018)]

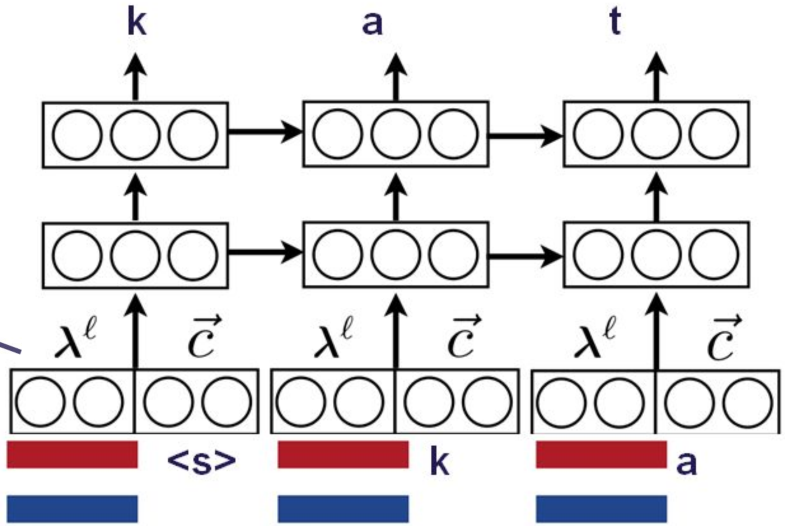
NLP Models *Encode* Typology

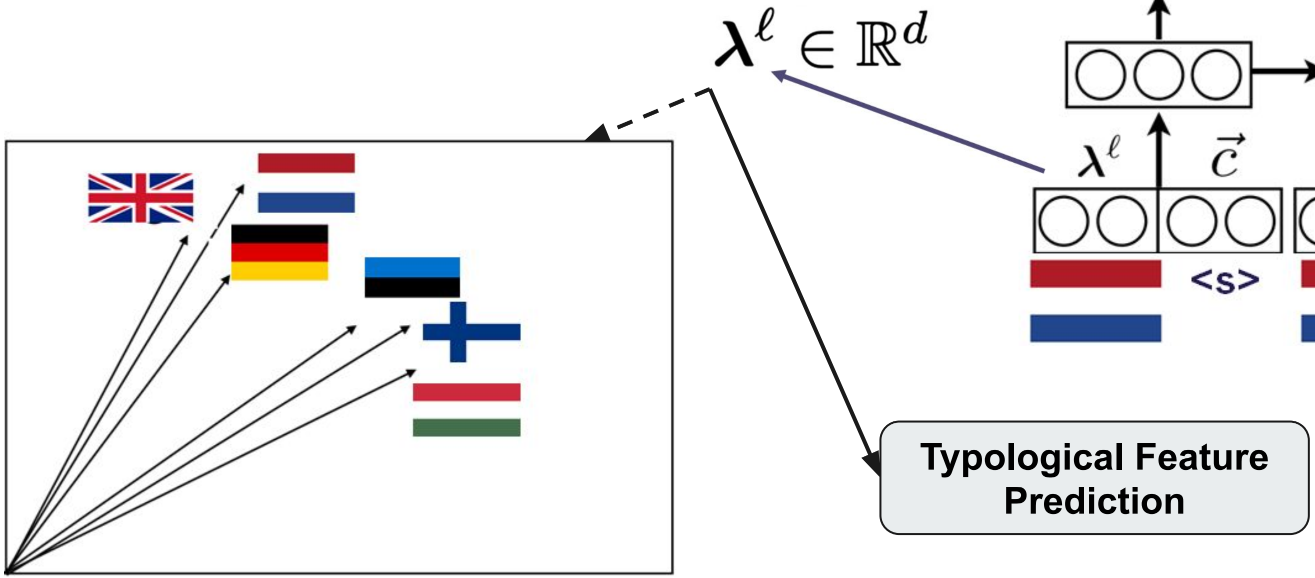
- ...or at least something that correlates strongly
- Typological Prediction from Language Representations
- Task-Dependent Change in Language Similarities



[Östling and Tiedemann, EACL 2017]

General Experimental Set-up





[Bjerva and Augenstein (NAACL 2018)]

Typological Prediction: Phonology

- 20 phonological features from WALS
- Grapheme-to-Phoneme
- 102 languages

```
(English, variation) -> ˌvɛəriˈeɪʃən
(French,  variation) -> ˌvaʁjaˈsjɔ̃
```

- A model trained on a **phonological task**, is better at predicting **phonological features**

System / features	Random phon.	Unseen phon.	All feat.
Most Frequent Class	*59.39%	63.71%	*58.12%
k-NN (fine-tuned)	53.09%	*77.45%	51.9%

[Bjerva and Augenstein (NAACL 2018)]

Typological Prediction: Morphology

- 41 morphological features from WALS
- SIGMORPHON -2017 Data
- 29 languages

```
(release, V;V.PTCP;PRS) -> releasing
```

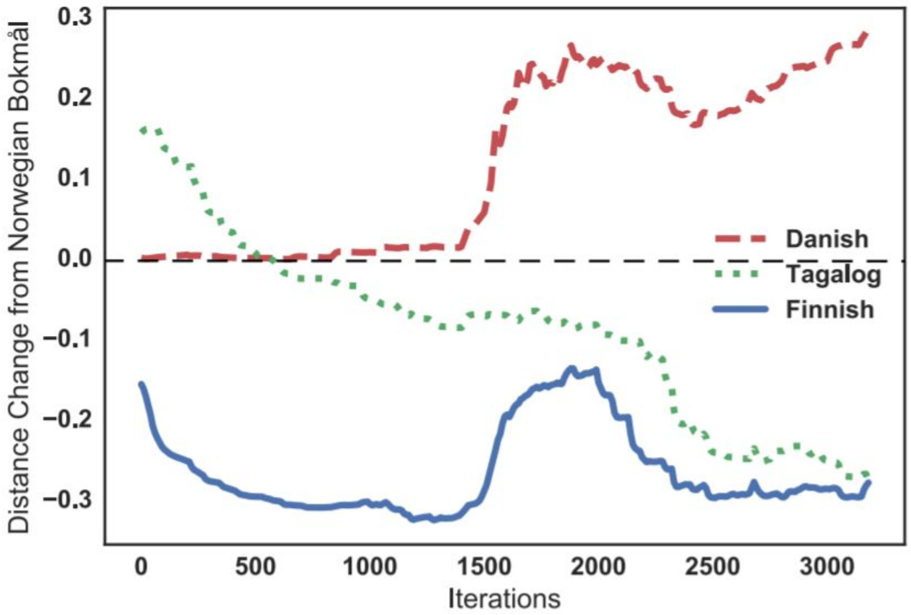
- A model trained on a **morphological task**, is better at predicting **morphological features**

System / Features	Random morph.	Unseen morph.	All feat.
Most Frequent Class	77.98%	85.68%	84.12%
k-NN (fine-tuned)	***82.91%**	***91.92%**	84.95%

[Bjerva and Augenstein (NAACL 2018),
Bjerva, Östling, Han Veiga, Tiedemann, Augenstein (CL 2019)]

Language Representations

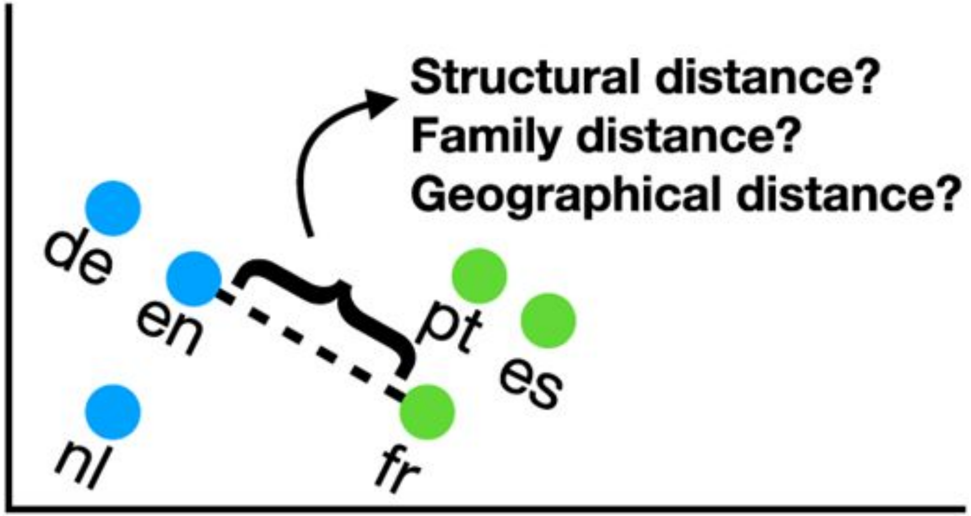
- Language similarities shift depending on the task



[Bjerva and Augenstein (NAACL 2018),
Bjerva, Östling, Han Veiga, Tiedemann, Augenstein (CL 2019)]

Language Representations

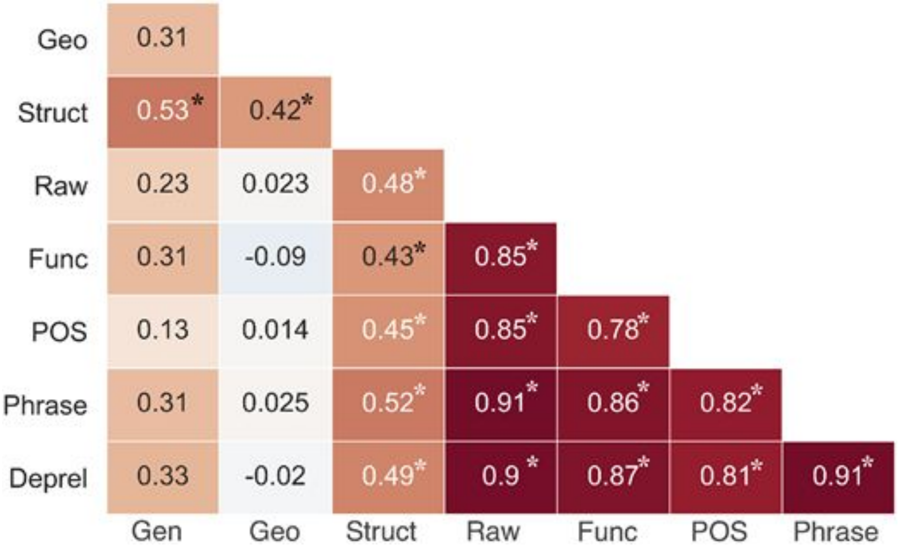
- Language similarities shift depending on the task
- What do Language Representations Represent?



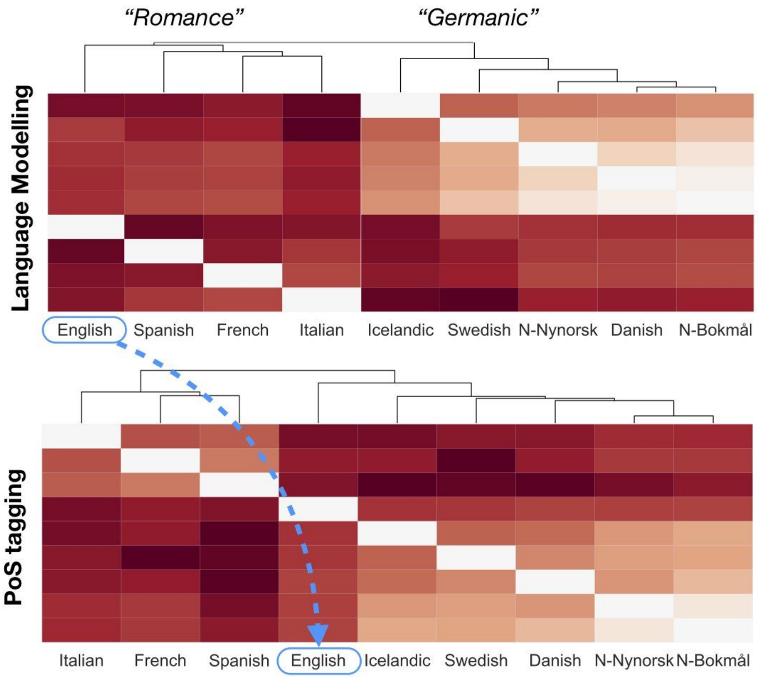
[Bjerva and Augenstein (NAACL 2018),
Bjerva, Östling, Han Veiga, Tiedemann, Augenstein (CL 2019)]

Language Representations

- Language similarities shift depending on the task
- What do Language Representations Represent?
- Generally seem to correspond to **structural similarities**



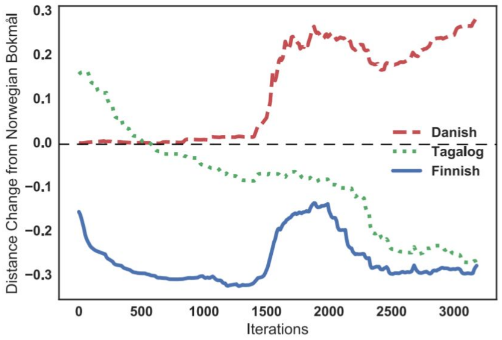
Language Similarities under Different Tasks



Summary: NLP Models Encode Typology

System / features	Random phon.	Unseen phon.	All feat.
Most Frequent Class	*59.39%	63.71%	*58.12%
k-NN (fine-tuned)	53.09%	*77.45%	51.9%

- Language representations can predict typology
  - Task-specific representations are better at related features
- Language representations evolve to make typologically similar languages more similar
  - Phonologically dissimilar languages diverge in a phonological task
  - Syntactically similar languages converge in a syntactic task



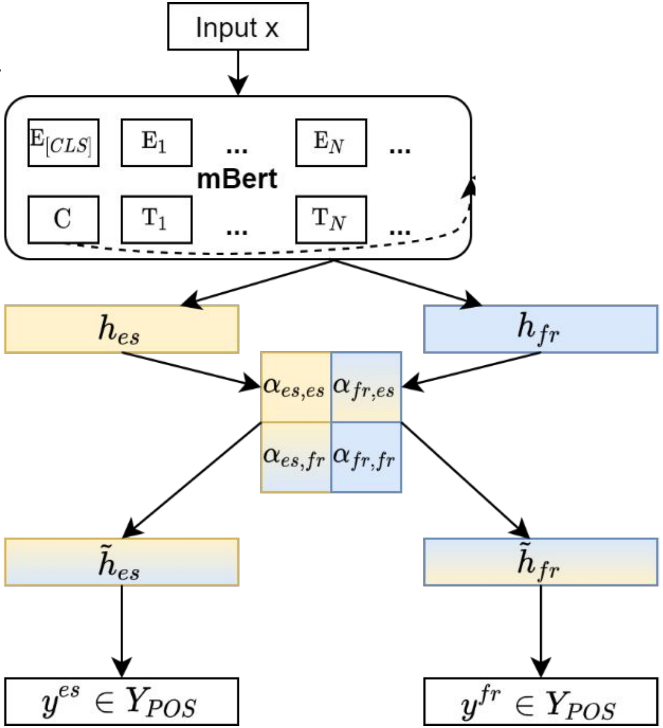
[Bjerva and Augenstein (EACL 2021)]

NLP Models *Need* Typology

- Can we force a model not to use typology?
  - Inspired by Domain Adversarial Networks [Ganin and Lepitsky, 2014] **-> blinding**
  - Using a "Sluice Network" [Ruder et al., AAAI 2019] **-> sharing**

- We introduce an auxiliary task based on **typological feature prediction**

- We add an *optional* gradient reversal **(-λ)**

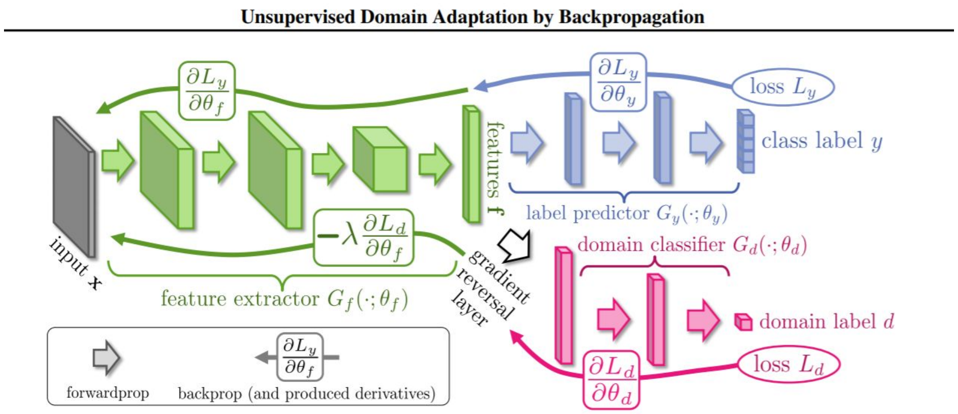


[Ganin and Lepitsky, 2014]

Gradient Reversal Layer / Typological Blinding

- Original formulation:
  - Domain-invariant features
  - Maximising loss on domain-prediction wrt. θ_f
  - Feature distributions as similar as possible, regardless of the domain

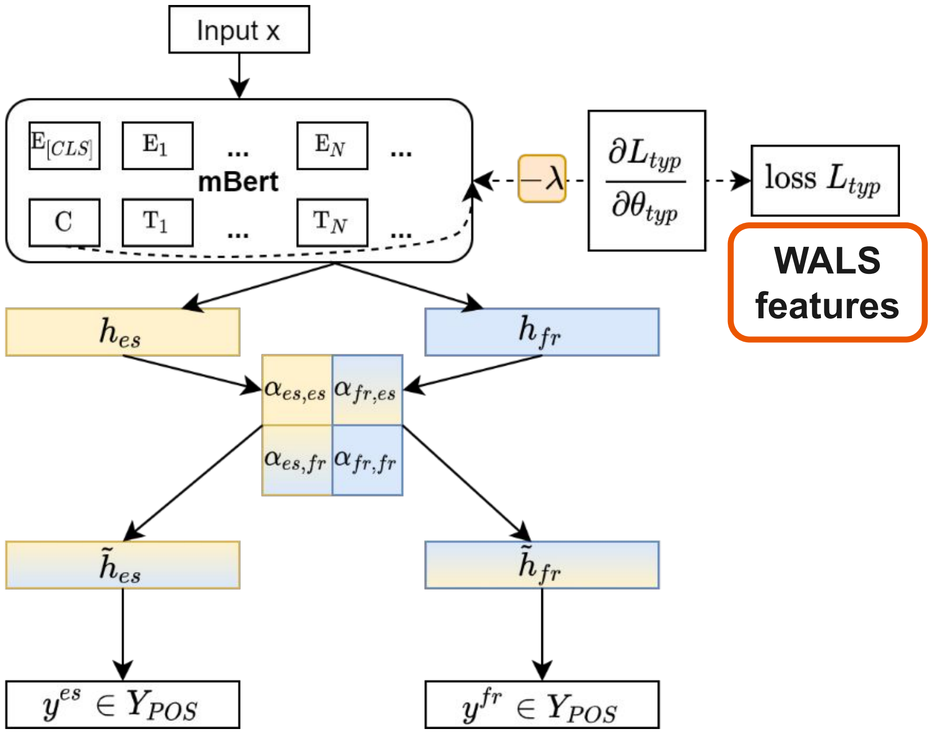
- Our formulation:
  - Feature distributions as similar as possible regardless of **typological feature**
  - Hypothesis: If the model relies on typology:
    - **Performance** will drop
    - **Sharing** will be affected



[Bjerva and Augenstein (EACL 2021)]

Experiments

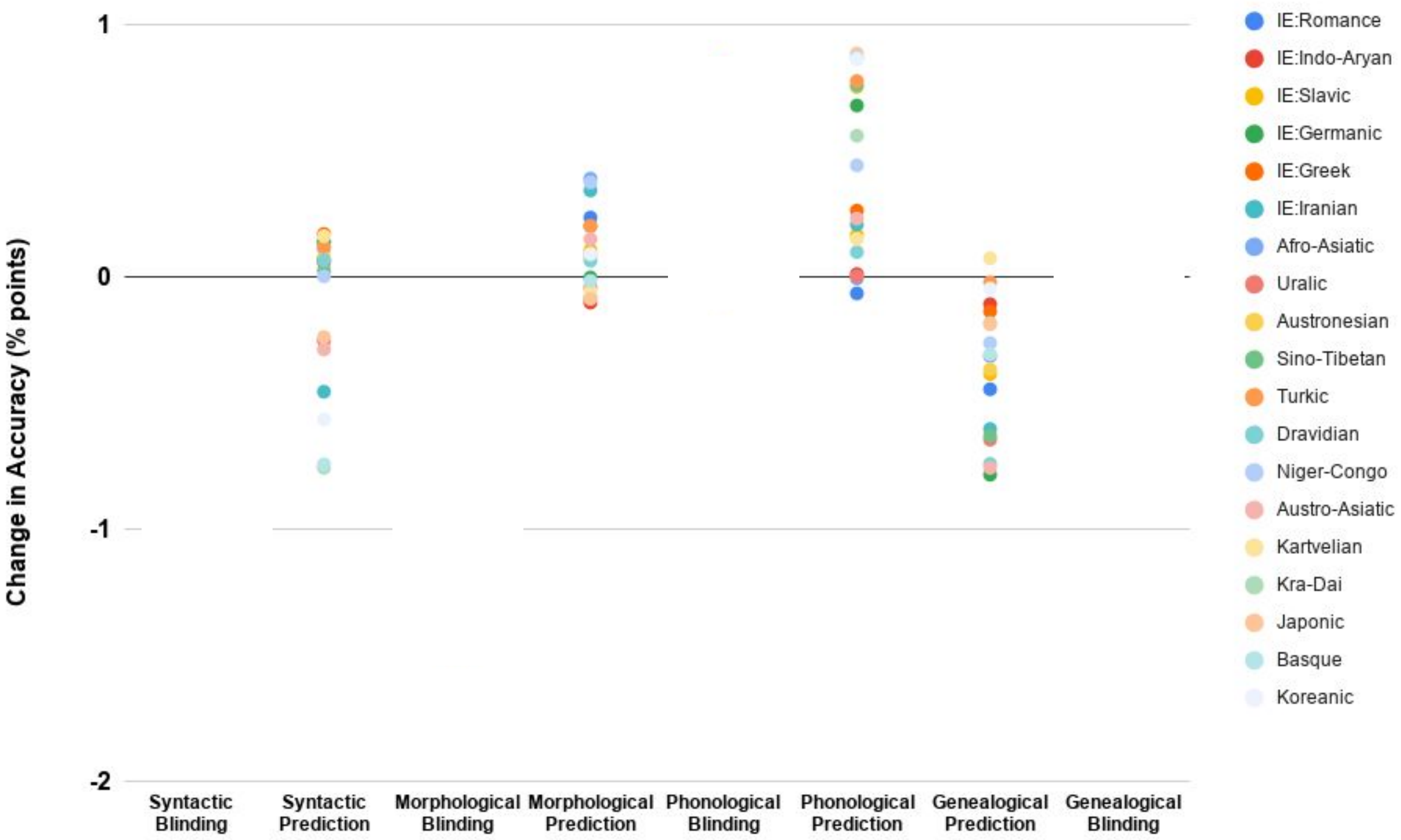
- Conditions:
  - **Blinding** vs. **Prediction**
  - Baseline: No blinding or prediction
- Typological Categories:
  - Syntax
  - Morphology
  - Phonology (**control**)
  - Geneaology (**control**)
- NLP Tasks:
  - Named Entity Recognition
  - POS tagging
  - Natural Language Inference
- Up to 40 languages (XTREME, Hu et al. 2020)



Research Questions

- How is the model's **performance** affected by blinding and prediction?
    - Does the type of feature blinded/predicted affect results in any particular way?
    - E.g. does blinding to a syntactic feature decrease performance on a task that requires syntax?

- How is the model's **sharing** affected by blinding/prediction?
    - Does the model represent syntactically similar languages differently under syntactic blinding?


1
0
-1
-2
Change in Accuracy (% points)
Syntactic Blinding
Syntactic Prediction
Morphological Blinding
Morphological Prediction
Phonological Blinding
Phonological Prediction
Genealogical Prediction
Genealogical Blinding
IE:Romance
IE:Indo-Aryan
IE:Slavic
IE:Germanic
IE:Greek
IE:Iranian
Afro-Asiatic
Uralic
Austronesian
Sino-Tibetan
Turkic
Dravidian
Niger-Congo
Austro-Asiatic
Kartvelian
Kra-Dai
Japonic
Basque
Koreanic

Research Questions

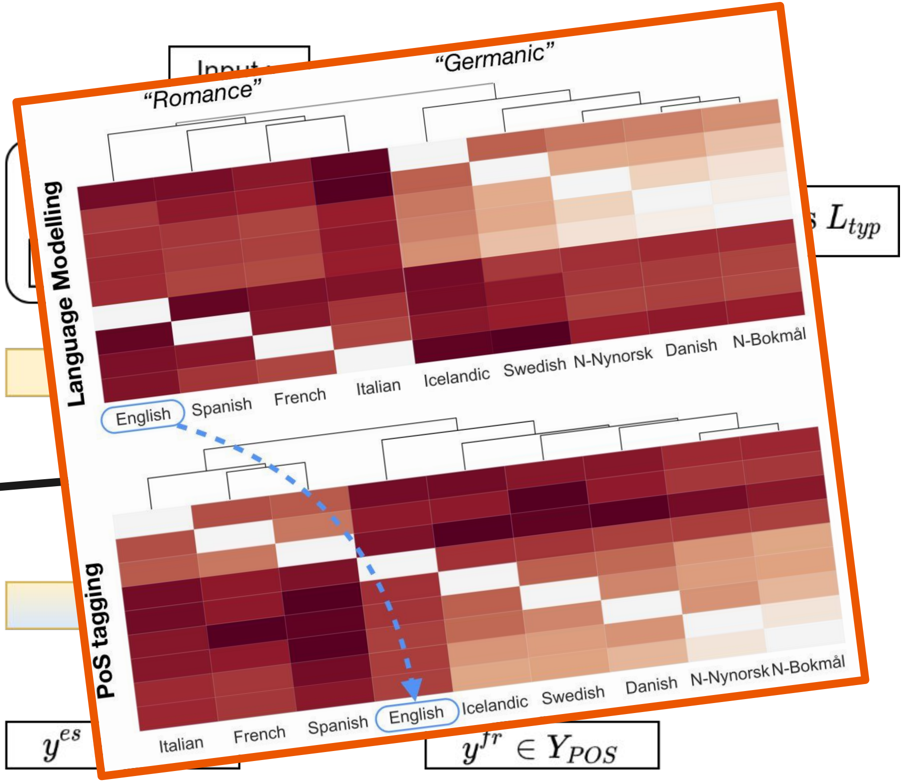
- How is the model's **performance** affected by blinding and prediction?
  - Does the type of feature blinded/predicted affect results in any particular way?
  - E.g. does blinding to a syntactic feature decrease performance on a task that requires syntax?
- How is the model's **sharing** affected by blinding/prediction?
  - Does the model represent syntactically similar languages differently under syntactic blinding?

[Bjerva and Augenstein (EACL 2021)]

Cross-lingual Sharing

Model	Struct.	Lang. Emb.
Syntactic Blind.	0.31	0.27
Morphological Blind.	0.34	0.29
Phonological Blind.	0.40	0.41
Genealogical Blind.	0.29	0.31
No blind./pred.	0.43	0.40
Syntactic Pred.	**0.52**	0.53
Morphological Pred.	0.49	**0.56**
Phonological Pred.	0.41	0.39
Genealogical Pred.	0.47	0.38

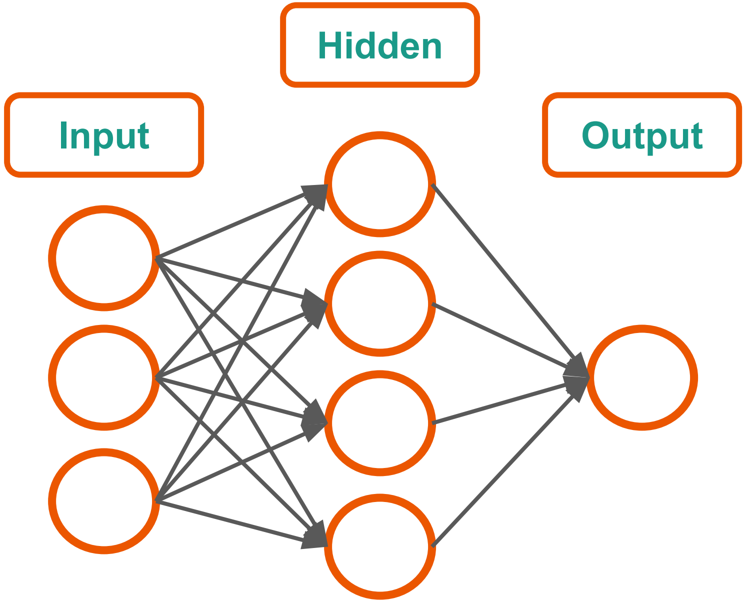
Table 2: Pearson correlations between α weights and language similarity measures.



So...

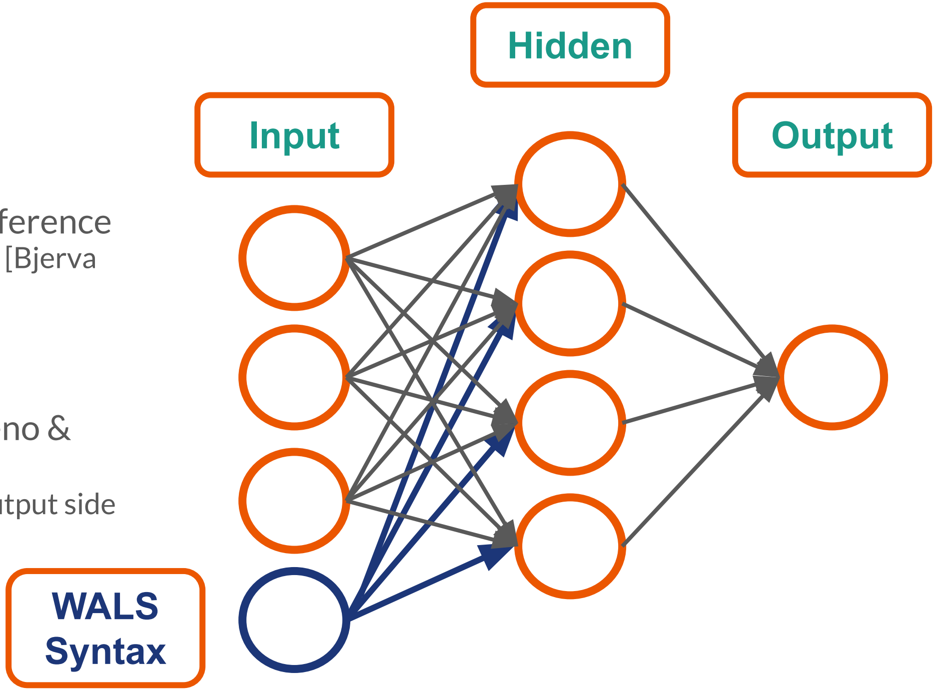
- NLP Models **Encode** Typology
- NLP Models **Need** Typology
- ...why doesn't it help to add more?

Incorporating Typological Knowledge



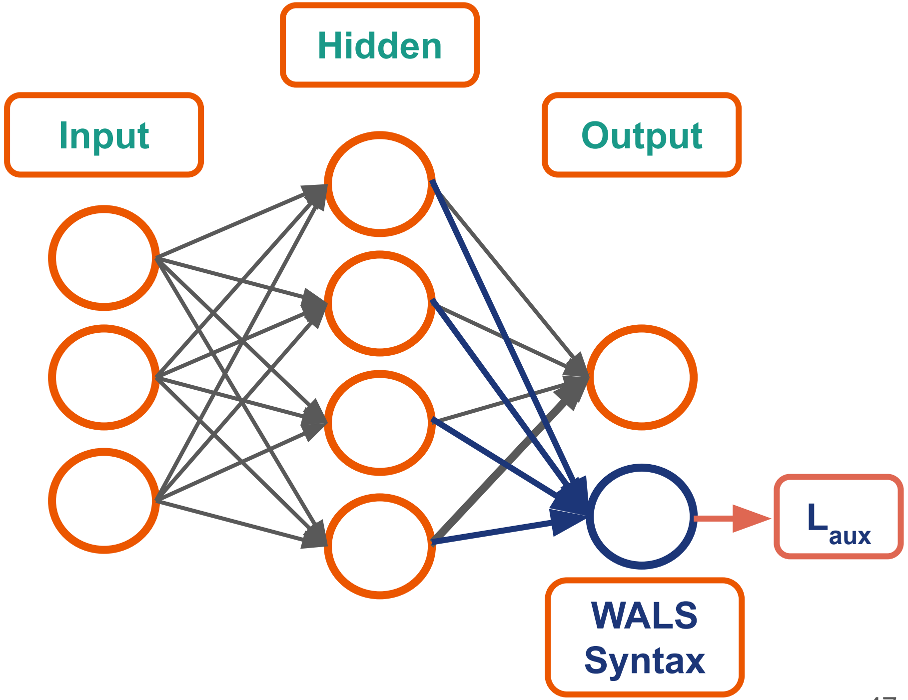
Incorporating Typological Knowledge

- As **features**
- Does not seem to make much of a difference
    - [Moreno & Oncevay, SIGTYP 2021], [Bjerva et al. (2018)]
    - See Miryam's keynote for parsing-experiments on this!
- Ongoing work, follow-up from [Moreno & Oncevay, SIGTYP 2021]
    - Typological features on input and output side of MT?



Incorporating Typological Knowledge

- As **auxiliary task**
- Does not seem to make much of a difference
  - [Bjerva and Augenstein, EACL 2021]



Incorporating Typological Knowledge

- Directly:
  - As **features**
  - As **auxiliary tasks**

- Indirectly:
  - Predict beneficial language pairings using typology
    (Dolicki and Spanakis, https://arxiv.org/pdf/2105.05975.pdf)
    - Results can be predicted, but do not seem to fit typological expectations
    - Best source language for Bulgarian:
      - French (POS)
      - Russian (NER)
      - Thai (NLI)

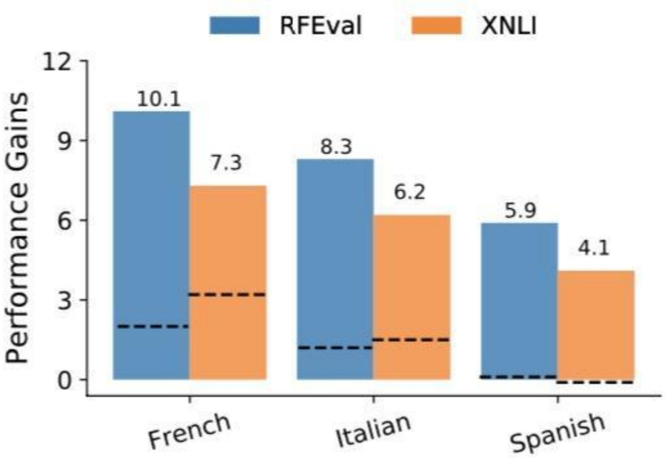
How can we inform models of Typology?

[*SEM 2021]

Is Typology the Solution?

she's eating a red apple

elle mange une rouge pomme



(b) Reversing adjective-noun order

Thanks to all of my collaborators!

[NAACL 2018]
From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings
Johannes Bjerva, **Isabelle Augenstein**

[CL 2019]
What do Language Representations Really Represent?
Johannes Bjerva, **Robert Östling, Maria H Veiga, Jörg Tiedemann, Isabelle Augenstein**

[NAACL 2019]
A Probabilistic Generative Model of Linguistic Typology
Johannes Bjerva, **Yova Kementchedjhieva, Ryan Cotterell, Isabelle Augenstein**

[SIGTYP 2020]
SIGTYP 2020 Shared Task: Prediction of Typological Features
Johannes Bjerva,
Elizabeth Salesky, Sabrina J Mielke, Aditi Chaudhary, Giuseppe GA Celano, Edoardo M Ponti, Ekaterina Vylomova, Ryan Cotterell, Isabelle Augenstein

[*SEM 2021]
Inducing language-agnostic multilingual representations
Wei Zhao, Steffen Eger, Johannes Bjerva, **Isabelle Augenstein**

[EACL 2021]
Does Typological Blinding Impede Cross-Lingual Sharing?
Johannes Bjerva, **Isabelle Augenstein**