

# Does Typological Blinding Impede Cross-Lingual Sharing?

Johannes Bjerva

Department of Computer Science  
Aalborg University  
jbjerva@cs.aau.dk

Isabelle Augenstein

Department of Computer Science  
University of Copenhagen  
augenstein@di.ku.dk

## Abstract

Bridging the performance gap between high- and low-resource languages has been the focus of much previous work. Typological features from databases such as the World Atlas of Language Structures (WALS) are a prime candidate for this, as such data exists even for very low-resource languages. However, previous work has only found minor benefits from using typological information. Our hypothesis is that a model trained in a cross-lingual setting will pick up on typological cues from the input data, thus overshadowing the utility of explicitly using such features. We verify this hypothesis by blinding a model to typological information, and investigate how cross-lingual sharing and performance is impacted. Our model is based on a cross-lingual architecture in which the latent weights governing the sharing between languages is learnt during training. We show that (i) preventing this model from exploiting typology severely reduces performance, while a control experiment reaffirms that (ii) encouraging sharing according to typology somewhat improves performance.

## 1 Introduction

Most languages in the world have little access to NLP technology due to data scarcity (Joshi et al., 2020). Nonetheless, high-quality multilingual representations can be obtained using only a raw text signal, e.g. via multilingual language modelling (Devlin et al., 2019). Furthermore, structural similarities of languages are to a large extent documented in typological databases such as the World Atlas of Language Structures (WALS, Dryer and Haspelmath (2013)). Hence, developing models which can take use typological similarities of languages is an important direction in order to alleviate language technology inequalities.

While previous work has attempted to use typological information to inform NLP models, our

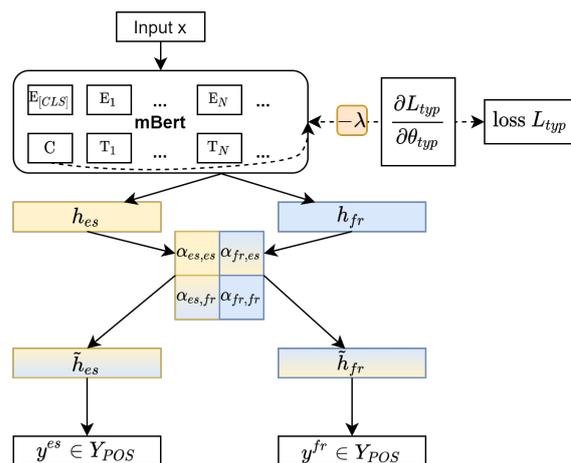


Figure 1: A PoS tagger is exposed (or blinded with gradient reversal,  $-\lambda$ ) to typological features. Observing  $\alpha$  values tells us how typology affects sharing.

work differs significantly from such efforts in that we *blind* a model to this information. Most previous work includes language information as features, by using language IDs, or language embeddings (e.g. Ammar et al. (2016); O’Horan et al. (2016); Östling and Tiedemann (2017); Ponti et al. (2019); Oncevay et al. (2020)). Notably, limited effects are usually observed from including typological features explicitly. For instance, de Lhoneux et al. (2018) observe positive cross-lingual sharing effects only in a handful of their settings. We therefore hypothesise that relevant typological information is learned as a by-product of cross-lingual training. Hence, although models do benefit from this information, it is not necessary to provide it explicitly in a high-resource scenario, where there is abundant training data. This is confirmed by Bjerva and Augenstein (2018a), who find that, e.g., language embeddings trained on a morphological task can encode morphological features from WALS.

In contrast with previous work, we *blind* a model to typological information, by using adversarial

techniques based on gradient reversal (Ganin and Lempitsky, 2014). We evaluate on the structured prediction and classification tasks in XTREME (Hu et al., 2020), yielding a total of 40 languages and 4 tasks. We show that when a model is blinded to typological signals relating to syntax and morphology, performance on related NLP tasks drops significantly. For instance, the mean accuracy across 40 languages for POS tagging drops by 1.8% when blinding the model to morphological features.

## 2 Model

An overview of the model is shown in Figure 1. We model each task in this paper using the following steps. First, contextual representations are extracted using multilingual BERT (m-BERT, Devlin et al. (2019)), a transformer-based model (Vaswani et al., 2017), trained with shared word-pieces across languages. We either blind m-BERT to typological features, with an added adversarial component based on gradient reversal (Ganin and Lempitsky, 2014), or expose it to them via multi-task learning (MTL, (Caruana, 1997)). Representations from m-BERT are fed to a latent multi-task architecture learning network (Ruder et al., 2019), which includes  $\alpha$  parameters we seek to investigate. The model learns which parameters to share between languages (e.g.  $\alpha_{es,fr}$  denotes sharing between Spanish and French).

### 2.1 Sharing architecture

Our sharing architecture is based on that of Ruder et al. (2019), which has latent variables learned during training, governing which layers and subspaces are shared between tasks, to what extent, as well as the relative weighting of different task losses. We are most interested in the parameters which control the sharing between the hidden layers allocated to each task, referred to as  $\alpha$  parameters (Ruder et al., 2019). Consider a setting with two tasks  $A$  and  $B$ . The outputs  $h_{A,k}$  and  $h_{B,k}$  of the  $k$ -th layer for task  $A$  and  $B$  interact through the  $\alpha$  parameters, for which the output is defined as:

$$\begin{bmatrix} \tilde{h}_{A,k} \\ \tilde{h}_{B,k} \end{bmatrix} = \begin{bmatrix} \alpha_{AA} & \alpha_{AB} \\ \alpha_{BA} & \alpha_{BB} \end{bmatrix} [h_{A,k}^\top, h_{B,k}^\top] \quad (1)$$

where  $\tilde{h}_{A,k}$  is a linear combination of the activations for task  $A$  at layer  $k$ , weighted with the learned  $\alpha$ s. While their model is an MTL model,

we choose to interpret this differently by considering each *language* as a task, yielding  $\alpha \in \mathbf{R}^{l \times l}$ , where  $l$  is the number of languages for the given task. Each activation  $\tilde{h}_{A,k}$  is then a linear combination of the language specific activations  $h_{A,k}$ . These are used for prediction in the downstream tasks, as in the baselines from Hu et al. (2020).

Crucially, this model allows us to draw conclusions about parameter sharing between languages by observing the  $\alpha$  parameters under the blinding and prediction conditions. We will combine this insight with observing downstream task performance in order to draw conclusions about the effects of typological feature blinding and prediction.

### 2.2 Blinding/Exposing a Model to Typology

We introduce a component which can either blind or expose the model to typological features. We implement this as a single task-specific layer per feature, using the [CLS] token from m-BERT model, without access to any of the soft sharing between languages from  $\alpha$ -layers. Each layer optimises a categorical cross-entropy loss function ( $L_{typ}$ ).

For this task, we predict typological features drawn from WALS (Dryer and Haspelmath, 2013), inspired by previous work (Bjerva and Augenstein, 2018a). Unlike previous work, we also blind the model to such features by including a gradient reversal layer (Ganin and Lempitsky, 2014), which multiplies the gradient of the typological prediction task with a negative constant ( $-\lambda$ ), inspired by previous work on adversarial learning (Goodfellow et al., 2014; Zhang et al., 2019; Chen et al., 2019). We hypothesise that using a gradient reversal layer for typology will yield typology-invariant features, and that this will perform worse on tasks for which the typological feature at hand is important. For instance, we expect that blinding a model to syntactic features will severely reduce performance for tasks which rely heavily on syntax, such as POS tagging.

## 3 Cross-Lingual Experiments

We investigate the effects of typological blinding, using typological parameters as presented in WALS (Dryer and Haspelmath, 2013). The experiments are run on XTREME (Hu et al., 2020), which includes up to 40 languages from 12 language families and two isolates. We experiment on the following languages (ISO 639-1 codes): af, ar, bg, bn, de, el, en, es, et, eu, fa, fi, fr, he, hi, hu, id, it, ja, jv, ka, kk, ko, ml, mr, ms, my, nl, pt, ru, sw,

ta, te, th, tl, tr, ur, vi, yo, and zh. We experiment on four tasks: POS (part of speech tagging), NER (named entity recognition), XNLI (cross-lingual natural language inference), and PAWS-X (paraphrase identification). Our general setup for the structured prediction tasks (POS and NER) is that we train on all available languages, and downsample to 1,000 samples per language. For the classification tasks XNLI and PAWS-X, we train on the English training data and fine-tune on the development sets, as no training data is available for other languages. Hence, typological differences will be the main factor in our results, rather than differences in dataset sizes.

### 3.1 Typological Prediction and Blinding

We first investigate whether prohibiting or allowing access to typological features has an effect on model performance using our architecture. We hypothesise that our multilingual model will leverage signals related to the linguistic nature of a task when optimising its its sharing parameters  $\alpha$ .

There exists a growing body of work on prediction of typological features (Daumé III and Campbell, 2007; Murawaki, 2017; Bjerva and Augenstein, 2018b; Bjerva et al., 2019a,b), most notably in a recent shared task on the subject (Bjerva et al., 2020). While we are inspired by this direction of research, our contribution is not concerned with the accuracy of the prediction of such features, and this is therefore not evaluated in detail in the paper.

Moreover, an increasing amount of work measures the correlation of predictive performance of cross-lingual models with typological features as a way of probing what a model has learned about typology (Malaviya et al., 2017; Choenni and Shutova, 2020; Gerz et al., 2018; Nooralahzadeh et al., 2020; Zhao et al., 2020). In contrast to such post-hoc approaches, our experimental setting allows for measuring the impact of typology on cross-lingual sharing performance in a direct manner as part of the model architecture.

**Syntactic Features** We first blind/expose the model to syntactic features from WALS (Dryer and Haspelmath, 2013). We take the set of word order features which are annotated for all languages in our experiments, resulting in 33 features. This includes features such as *81A: Order of Subject, Object and Verb*, which encodes what the preferred word ordering is (if any) in a transitive clause. For all features, we exclude feature values which do

not occur for our set of languages. We hypothesise that performance will drop for all four tasks, as they all require syntactic understanding.

**Morphological Features** We next attempt to blind/expose the model to the morphological features in WALS. We use the same approach as above, resulting in a total of 8 morphological features. This includes features such as *26A: Prefixing vs. Suffixing in Inflectional Morphology*, indicating to what extent a language uses prefixing or suffixing morphology. We hypothesise that mainly the POS tagging task will suffer under this condition, whereas other tasks only to some extent require morphology.

**Phonological Features** We next consider a control experiment, in which we attempt to blind/expose the model to phonological features in WALS. We arrive at a total of 15 phonological features, such as *1A: Consonant Inventories* which indicates the size of the consonant inventory of a language. We expect the performance to remain relatively unaffected by this task, as phonology ought to have little importance given a textual input.

**Genealogical Features** Finally, we attempt to use what one might consider to be language metadata. We attempt to blind/expose the model to what language family a language belongs to. This can be seen as a type of proxy to language similarity, and correlates relatively strongly with structural similarities in languages. Because of this correlation with structural similarities, we expect blinding under this condition to only slightly reduce performance for all tasks, as previous work has shown this type of relationship not to be central in language representations (Bjerva et al., 2019c).

### 3.2 Results

In general, we observe a drop in performance when blinding the model to relevant typological information, and an increase in performance when exposing the model to it (Table 1). For phonological blinding or prediction, none of the four tasks is noticeably affected. Although, e.g., both the syntactic and morphological prediction tasks increase performance on POS tagging, it is not straightforward to draw conclusions on which of these is the most efficient, as there is a substantial correlation between syntactic and morphological features. As for XNLI and PAWS-X, performance notably drops under both the syntactic and genealogical blinding tasks.

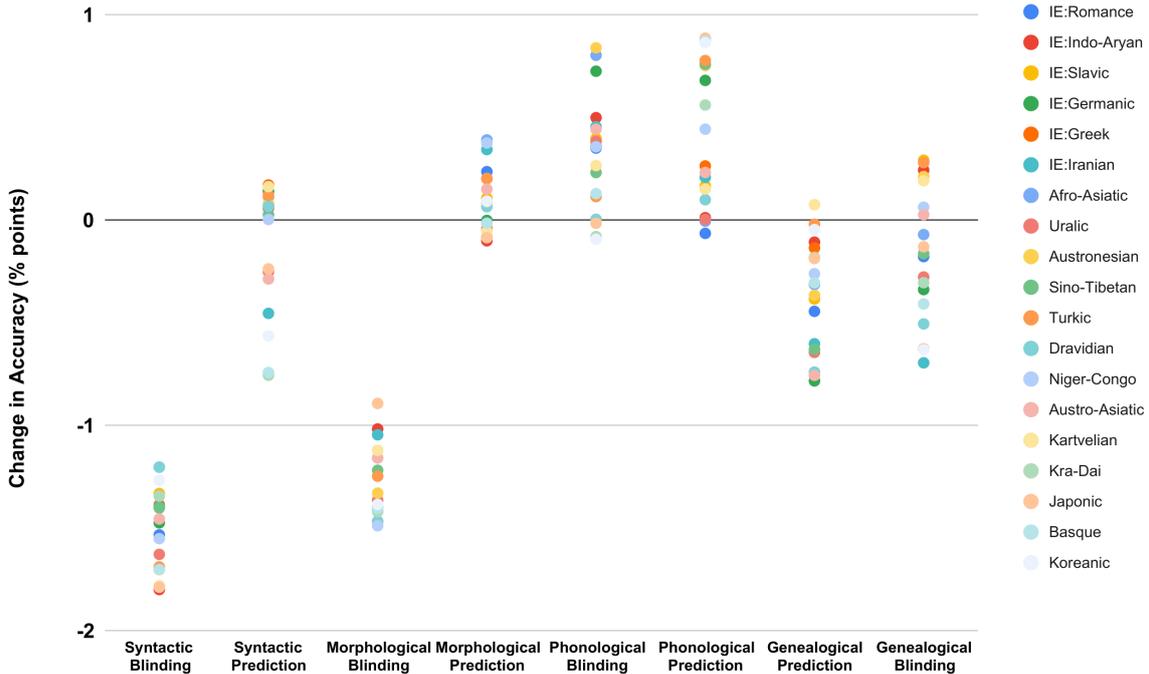


Figure 2: PoS tagging results per language family across blinding and prediction conditions

Model	POS	NER	XNLI	PAWS-X
+ Syntactic Blind.	85.3 <sup>-</sup>	76.4	64.2 <sup>-</sup>	80.6 <sup>-</sup>
+ Morphological Blind.	85.0 <sup>-</sup>	77.2	64.9	81.4
+ Phonological Blind.	86.7	77.1	65.0	81.6
+ Genealogical Blind.	86.1	77.0	64.7	81.1
<i>m-BERT baseline</i>	86.8	77.3	65.1	81.7
+ Syntactic Pred.	87.0	77.5	<b>65.3<sup>+</sup></b>	<b>81.9<sup>+</sup></b>
+ Morphological Pred.	<b>87.2<sup>+</sup></b>	77.3	65.2	81.7
+ Phonological Pred.	86.7	77.1	65.0	81.7
+ Genealogical Pred.	87.0	<b>77.6</b>	<b>65.3<sup>+</sup></b>	81.8

Table 1: Typological Blinding and Prediction. Mean POS accuracy, NER F1 scores, XNLI accuracy and PAWS-X accuracy across all languages. + and - indicate significantly better or worse performance respectively, as determined by a one-tailed t-test ( $p < 0.01$ ).

Figure 2 shows results for PoS tagging under prediction and blinding across language families, following the same scheme as Hu et al. (2020). Interestingly, the syntactic and morphological blinding settings are robust across all language families, yielding a drop in accuracy across the board. All other conditions yield mixed results. This further strengthens our argument that preventing a model from learning syntactic and morphological features can be severely detrimental.

### 3.3 The Effect of Typology on Latent Architecture Learning

The results show that preventing access to typological features hampers performance, whereas providing access improves performance. We now turn to an analysis of how the model shares parameters across languages in this setting. Our hypothesis is that blinding will prevent models from sharing parameters between similar languages, in spite of typological similarities. Concretely, we expect that the drop in POS tagging performance under morphological blinding is caused by lower  $\alpha$  weights between languages which are morphologically similar, and higher  $\alpha$  weights between languages which are dissimilar. Recall that these parameters are latent variables learned by the model, regulating the amount of sharing between languages (see Eq. 1). We investigate the correlations between the  $\alpha$  sharing parameters, and two proxies of language similarity. We focus on the POS task, as the results from the typological blinding and prediction experiments were the most pronounced here, as both morphological and syntactic blinding affected performance.

Our first measure of language similarity is based on Bjerva et al. (2019c), who introduce what they refer to as structural similarity. This is based on

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.	<b>0.52</b>	0.53
Morphological Pred.	0.49	<b>0.56</b>
Phonological Pred.	0.41	0.39
Genealogical Pred.	0.47	0.38

Table 2: Pearson correlations between  $\alpha$  weights and language similarity measures.

dependency statistics from the Universal Dependencies treebank (Zeman et al., 2020), resulting in vectors which describe how different syntactic relations are used in each language. Previous work has shown that this measure of similarity correlates strongly with that learned in embedded language spaces during multilingual training. In addition to considering these dependency statistics, we also use language embeddings drawn from Östling and Tiedemann (2017). For each language similarity measure we calculate its pairwise Pearson correlation with the  $\alpha$  values learned under each condition.

Table 2 shows correlations between  $\alpha$  weights and similarities increase when predicting typological features, and decreases when blinded to such features. Hence, when the model has indirect access to, e.g., the SVO word ordering features of languages, sharing also reflects this.

## 4 Discussion

We have shown that blinding a multilingual model to typological features severely affects sharing across a relatively large language sample, and for several NLP tasks. The effects on model performance, as evaluated over 40 languages and 4 tasks from XTREME (Hu et al., 2020), were the largest for POS tagging. The fact that smaller effects were observed for NER, could be because this task relies more on memorising NEs rather than using (morpho-)syntactic cues (Augenstein et al., 2017). Furthermore, the relatively small effects on XNLI and PAWS-X can also be interpreted as evidence for that typology is less important in these tasks than in more traditional linguistic analysis.

A potential critique of our approach is that it merely blinds the model to language identities.

This could be the case, if only some latent representation of, e.g., “SVO” ordering is used to represent a language identity. However, previous work has shown that morphological information is encoded by the type of model we investigate. Hence, since we only blind features in a single category at a time, we expect that the model’s representation of language identities is unaffected.

Not only do we observe a drop in performance when blinding a model to syntactic features, but we also observe that the  $\alpha$  sharing weights in our model do not appear to correlate with linguistic similarities in this setting. Conversely, encouraging a model to consider typology, by jointly optimising it for typological feature prediction, improves performance in general. Furthermore,  $\alpha$  weights in this scenario converge towards correlating with structural similarities of languages. This is in line with recent work which has found that m-BERT uses fine-grained syntactic distinctions in its cross-lingual representation space (Chi et al., 2020).

We interpret this as evidence for the fact that typology can be a necessity for modelling in NLP. Our results furthermore corroborate previous work in that we only find moderate benefits from including typological information explicitly. We expect that this to a large degree is due to the typological similarities of languages being encoded implicitly based on correlations between patterns in the input data. As low-resource languages often do not even have access to any substantial amount of raw text, but often do have annotations in WALS, we expect that using typological information can go some way towards building truly language-universal models.

## 5 Conclusions

We have shown that preventing access to typology can impede the performance of cross-lingual sharing models. Investigating latent weights governing the sharing between languages shows that this prevents the model from sharing between typologically similar languages, which is otherwise learned based on patterns in the input. We therefore expect that using typological information can be of particular interest for building truly language-universal models for low-resource languages.

## Acknowledgements

This research has received funding from the Swedish Research Council (grant No 2019-04129), and the NVIDIA Corporation (Titan Xp GPU).

## References

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.
- Johannes Bjerva and Isabelle Augenstein. 2018a. [From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.
- Johannes Bjerva and Isabelle Augenstein. 2018b. [Tracking Typological Traits of Uralic Languages in Distributed Language Representations](#). In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.
- Johannes Bjerva, Yova Kementchedjheva, Ryan Cotterell, and Isabelle Augenstein. 2019a. [A probabilistic generative model of linguistic typology](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.
- Johannes Bjerva, Yova Kementchedjheva, Ryan Cotterell, and Isabelle Augenstein. 2019b. [Uncovering probabilistic implications in typological knowledge bases](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.
- Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Celano Giuseppe, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. [SIGTYP 2020 shared task: Prediction of typological features](#). In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.
- Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019c. [What Do Language Representations Really Represent?](#) *Computational Linguistics*, 45(2):381–389.
- Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28 (1):41–75.
- Steven Chen, Nicholas Carlini, and David Wagner. 2019. Stateful detection of black-box adversarial attacks. *arXiv preprint arXiv:1907.05587*.
- Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. *arXiv preprint arXiv:2005.04511*.
- Rochelle Choenni and Ekaterina Shutova. 2020. [What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties](#). *CoRR*, abs/2009.12862.
- Hal Daumé III and Lyle Campbell. 2007. [A Bayesian Model for Discovering Typological Implications](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.
- Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. [On the relation between linguistic typology and \(limitations of\) multilingual language modeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.
- Pratik Joshi, Sebastian Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.
- Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997.

- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. [Learning language representations for typology prediction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.
- Yugo Murawaki. 2017. [Diachrony-aware induction of binary latent representations from typological features](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-Shot Cross-Lingual Transfer with Meta Learning. In *Proceedings of EMNLP*. Association for Computational Linguistics.
- Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.
- Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of EMNLP*. Association for Computational Linguistics. ArXiv preprint arXiv:2004.14923.
- Robert Östling and Jörg Tiedemann. 2017. [Continuous multilinguality with language vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.
- Edoardo Maria Ponti, Helen O’horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Daniel Zeman, Joakim Nivre, and Mitchell Abrams et al. 2020. [Universal dependencies 2.6](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. 2019. The limitations of adversarial training and the blind-spot attack. *arXiv preprint arXiv:1901.04684*.
- Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing Language-Agnostic Multilingual Representations. *arXiv preprint arXiv:2008.09112*.