Uncovering Probabilistic Implications in Typological Knowledge Bases

Johannes Bjerva[ə] **Yova Kementchedjhieva**[ə] **Ryan Cotterell**[?,fi] **Isabelle Augenstein**[ə]

[ə]Department of Computer Science, University of Copenhagen
[?]Department of Computer Science, Johns Hopkins University
[fi]Department of Computer Science and Technology, University of Cambridge
`bjerva,yova,augenstein@di.ku.dk, rdc42@cam.ac.uk`

Abstract

The study of linguistic typology is rooted in the implications we find between linguistic features, such as the fact that languages with object-verb word ordering tend to have postpositions. Uncovering such implications typically amounts to time-consuming manual processing by trained and experienced linguists, which potentially leaves key linguistic universals unexplored. In this paper, we present a computational model which successfully identifies known universals, including Greenberg universals, but also uncovers new ones, worthy of further linguistic investigation. Our approach outperforms baselines previously used for this problem, as well as a strong baseline from knowledge base population.

1 Introduction

Linguistic typology is concerned with mapping out the relationships between languages with reference to structural and functional properties (Croft, 2002). A typologist may ask, for instance, how a language encodes syntactic features and relationships. Does it place its verbs before objects or after, and does it have prepositions or postpositions? It is well established that many features of languages are highly correlated, sometimes to the extent that they imply each other. Based on this observation, Greenberg (1963) establishes the notion of implicational universals, i.e., cases where the presence of one feature *strictly* implies the presence of another.

Universals are important to investigate as they offer insight into the inner workings of language and define the space of plausible languages. Universals can aid cognitive scientists examining the underlying processes of language, as there arguably is a cognitive reason for why, e.g., languages with OV ordering are postpositional (Greenberg, 1963). In the context of natural language processing (NLP), when creating synthetic data for multilingual NLP,
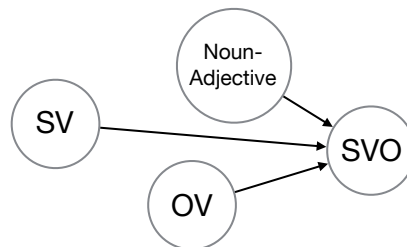

Figure 1: Visualisation of a section of our induced graphical model. Observing the features in the leftmost nodes (SV, OV, and Noun-Adjective), can we correctly infer the value of the right-most node (SVO)?

one should consider universals to maintain the plausibility of the data (Wang and Eisner, 2016). Computational typology can furthermore be used to induce language representations, useful in, e.g., language modelling (Östling and Tiedemann, 2017) and syntactic parsing (de Lhoneux et al., 2018).

In this paper, we argue that the deterministic Greenbergian view of implications (Greenberg, 1963) is outdated. Instead, we suggest that a probabilistic view of implications is more suitable, and define the notion of a probabilistic typological implication as a certain conditional probability distribution. We do this by first placing a joint distribution over the vector of typological features, and then marginalising out all features other than the two under consideration. This computation is made tractable by learning a tree-structured graphical model (Figure 1) with the PC algorithm of Neapolitan (2004) and then applying the belief propagation (BP) algorithm (Pearl, 1982). We draw inspiration from manual linguistic efforts to this problem (Greenberg, 1963; Lehmann, 1978), as well as from previous computational methods (Daumé III and Campbell, 2007; Bjerva et al., 2019a). Additionally, we provide a qualitative analysis of predicted implications, as well as performing an empirical evaluation on typological feature prediction, com-

paring to strong baselines.

2 From A Generative Model to Probabilistic Implications

Notation. We now seek a probabilistic formalisation of typological implications. First, we will introduce the relevant notation. Let ℓ be a language. We will seek to explain the observed, language-specific binary vector of typological features, or parameters, $\boldsymbol{\pi}^\ell$ where $\pi_i^\ell = 1$ indicates that the ith typological feature is "on" in language ℓ. When it is unambiguous, we will drop the superscript ℓ. Note that we call the vector $\boldsymbol{\pi}$ due to a spiritual similarity to the principle-and-parameters framework of Chomsky (1981).

A Generative Model of Typology. We construct a simple generative probability model over the the vector of typological features $\boldsymbol{\pi}$, which factorises according to some tree structure $\mathcal{T}$. We will discuss the provenance of $\mathcal{T}$ below. Concretely, this distribution is defined as

$$p(\boldsymbol{\pi}) = \prod_{i=1}^{|\boldsymbol{\pi}|} p(\pi_i \mid pa_{\mathcal{T}}[\pi_i]) \qquad (1)$$

where $pa_{\mathcal{T}}[\cdot]$ is a function that returns the parents of π_i, if any, in the tree $\mathcal{T}$. Each conditional $p(\pi_i \mid pa_{\mathcal{T}}[\pi_i])$ is treated tabularly with one parameter per table entry: each table entry is a unique configuration of the feature π_i and its parents $pa_{\mathcal{T}}[\pi_i]$. We place a symmetric Dirichlet prior with concentration parameter $\alpha = 5$, over each of $p(\pi_i \mid pa_{\mathcal{T}}[\pi_i])$'s table entries. This corresponds to add-5 smoothing.

Probabilistic Implications. Although the original Greenbergian view of implications is deterministic, we argue that a probabilistic approach is more suitable. Indeed, logical implications are a special case of conditional probabilities that only take the values 0 and 1, rather than values in $[0, 1]$. Specifically, we argue that probabilistic implications should take the form of the following conditional probability distribution:

$$p(\pi_i \mid \pi_j) = \sum_{\boldsymbol{\pi}'} p(\pi_i, \boldsymbol{\pi}' \mid \pi_j) \qquad (2)$$

where $\boldsymbol{\pi}'$ is a subvector that omits the indices i and j. In text, our goal is to sum out all possible languages, holding two typological features, π_i and π_j, fixed. We note that since our model p

factorises according to the tree $\mathcal{T}$, this sum may be performed in polynomial time using dynamic programming, specifically the belief propagation algorithm (Pearl, 1982). Note that we contend this improves upon the ideas of Daumé III and Campbell (2007), who only considered pair-wise interactions of features: Our definition of probabilistic implications *marginalises out* all other features.

Discovering Probabilistic Implications. How can we use a generative model to discover typological implications? What we would like to know is how often $p(\pi_i \mid \pi_j)$ is significantly different than $p(\pi_i)$. We note that $p(\pi_i)$ can also be computed with BP. We now reduce the search for typological implications as asking when the quantity $|p(\pi_i \mid \pi_j) - p(\pi_i)|$ is statistically significantly greater than 0. Given a sufficiently expressive generative model p, this allows for a richer notion of implication than Greenberg original proposed, as it admits the softer notion of typological influence.

Learning the Structure of p. There are many ways to learn the tree structure $\mathcal{T}$, and we choose the PC algorithm of Neapolitan (2004). This algorithm works in two steps—first, it learns a skeleton graph from the data (in our case, a typological data base), with *undirected* edges. Next, it orients these edges so as to form a directed acyclic graph. Once we have fit this graph so as to represent $p(\boldsymbol{\pi})$, we are left with a tractable model we can use to predict held-out typological features and discover typological implications.

Parameter Estimation. We apply maximum a posteriori (MAP) inference in order to estimate the parameters of our model. If all the data were observed, i.e. there were no missing values in WALS, this could be achieved by counting and normalising across the typological database in question with the previously mentioned Dirichlet prior. (The prior simply corresponds to add-λ smoothing.) However, in many cases we do have missing data. In fact, we almost never observe all the values in WALS. Thus, we must rely on expectation-maximisation to perform MAP estimation (Dempster et al., 1977). The gist of the algorithm is simple: we compute "pseudocounts" for the missing entries using belief propagation, which we smooth as if they had been observed values. Using these pseudocounts, we get a new estimate of the parameters by count-and-divide as in the fully supervised case. We iterate between updating the pseudocounts and perform-
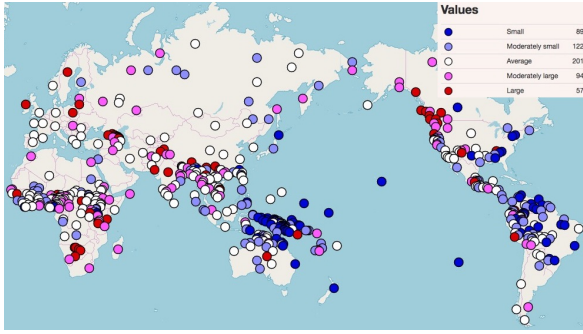
Figure 2: Consonant inventory sizes across languages in the world (WALS, Dryer and Haspelmath (2013)).

N implicants	2	3	4	5	6
Phonology	0.75	0.82	0.84	0.86	**0.89**
Morphology	0.77	0.85	**0.87**	0.70	0.82
Nominal Categories	0.72	0.83	0.80	**0.84**	0.81
Nominal Syntax	0.77	**0.89**	0.85	**0.89**	0.81
Verbal Categories	0.80	0.84	0.80	0.86	**0.90**
Word Order	0.74	0.86	0.86	0.86	**0.93**
Clause	0.75	0.81	0.84	**0.85**	0.84
Complex	0.82	0.83	0.87	**0.93**	0.84
Lexical	0.83	0.76	0.75	**0.85**	0.79
Mean	0.77	0.83	0.83	**0.85**	**0.85**
Most freq.			*0.30*		
Pairwise			*0.77*		
PRA			*0.81*		
Language embeddings			*0.85*		

Table 1: Accuracies for feature prediction in a typologically diverse test set, across number of implicants used. Note that the numbers are not comparable across columns nor to the baseline, since each makes a different number of predictions.

ing count-and-divide. This is a standard technique in the literature.

Decoding. In section 4, we are interested in predicting typological features given others. If we wish to predict π_i given observed features for a language $\boldsymbol{\pi}_{obs}$, we compute

$$\pi_i^\star = \operatorname{argmax} p(\pi_i \mid \boldsymbol{\pi}_{obs}) \quad (3)$$
$$= \operatorname{argmax} \sum_{\boldsymbol{\pi}_{unobs}} p(\pi_i, \boldsymbol{\pi}_{unobs} \mid \boldsymbol{\pi}_{obs}) \quad (4)$$

where we marginalize out all those features $\boldsymbol{\pi}_{unobs}$ unobserved or held out in a given language. The conditional may be computed with belief propagation and the argmax is over the set $\{0, 1\}$. This makes the computation tractable.

3 WALS: A Typological Database

Before explaining our experimental setup, we first explain the data set we use in evaluation. We evaluate on the World Atlas of Language Structure (WALS, Dryer and Haspelmath (2013), which is the largest openly available typological database. It comprises approximately 200 linguistic features with annotations for more than 2,500 languages. These annotations have been made by expert typologists through meticulous study of grammars and field work. WALS is quite sparse, however, as only 100 of these languages have annotations for all features. For instance, Figure 2 shows the distribution of consonant inventory sizes across the languages for which this feature is annotated. Although this is not our main contribution, the fact that we can predict held-out features offers a way to fill in the feature value gaps which exist for the vast majority of languages.

Pre-processing We pre-process our data similarly to Daumé III and Campbell (2007). We fil-

ter out features which are not encoded for at least 100 languages, and feature values which occur for fewer than 10% of the languages. The reason for this is that any implications found for exceedingly rare features is likely to be inconclusive. We further follow Daumé III and Campbell (2007) in that we binarise features with more than 7 feature values such that they simply encode whether or not a language has a feature. For instance, features are not likely to have implicants determining the *number* of tones, but rather the presence or absence of tones. Finally, they take into account that languages are not independent, as phylogenetic similarity can help infer features in closely related languages. We do not use this information, as we are interested in finding implications which ought to be independent of language relatedness.

4 Two Typological Experiments

In order to evaluate our probabilistic approach to typological implications, we define two tasks. Our **empirical evaluation** is based on predicting features so as to get an objective measure of our model, which is comparable both to previous work and other strong baselines. Second, we include a **qualitative evaluation**, as we are interested in uncovering both known and novel typological implications.

4.1 Predicting Typological Features

Feature prediction is a commonly used task in evaluating how well a given model is able to explain

the typological features of languages (Daumé III and Campbell, 2007; Malaviya et al., 2017; Cotterell and Eisner, 2018; Ponti et al., 2018; Bjerva et al., 2019a). This is an important task which can highlight the extent to which a model has learned interdependencies between languages and features. We include this evaluation to first show that our model has predictive power which surpasses strong baselines, before investigating the main research question of this work, i.e., the extent to which we can uncover probabilistic implications. We evaluate the models on feature prediction by fitting our model on 80% of the languages in WALS, and leaving out 10% of the languages for development and testing, respectively.

We split our evaluation of our model up across the feature categories present in WALS. These cover areas such as phonology, morphology etc., listed in Table 1. During the typological feature prediction experiments, we consider a single such WALS category at a time. We vary the number of implicants by allowing the model to observe 2 to 6 features from within this category as well as the values of features in *other* categories. This is done as having access to, e.g., all word-order features when predicting a final word order feature would be much easier than our setting. Hence, our experiment will show the extent to which increasing the number of features from the current feature category affects predictive power. We vary the number of implicants k from 2 to 6 features in each category with a total of n features, this gives us $\binom{n}{k}$ total sets per number of implicants k. For each such set, we attempt to predict all held-out features in that category in a leave-one-out-style evaluation. This results in $\binom{n}{k}(n-k)$ predictions to make per category per number of implicants k. Performance is measured by averaging the accuracy of predictions of all held-out features over the entire test set, across categories.

Baseline #1: Most frequent Since many typological features have low-entropy distributions, a most frequent class baseline is a relatively strong lower bound for prediction of typological features. For instance, this yields an accuracy of 45% when predicting the canonical subject–object–verb ordering in a language.

Baseline #2: Pairwise prediction We implement a simple baseline based on pairwise prediction of typological features. This is inspired by the

approach in Daumé III and Campbell (2007). As this code was not publicly available we provide our own non-Bayesian implementation.

Baseline #3: PRA Since WALS can be seen as a knowledge base, we apply a strong baseline from the field of knowledge base population. Path Ranking Algorithm (PRA) is an algorithm which finds relation paths by traversing the knowledge graph, which can then be used to predict implicatures and feature values (Lao and Cohen, 2010; Lao et al., 2011).[1] We train PRA using the standard hyperparameters of the existing implementation, which includes regularising with $\ell_1 = 0.001$ and $\ell_2 = 0.001$, as well as using negative sampling.

Baseline #4: Language embeddings Although we aim to predict *implications*, and not only feature values, we compare with previous work on predicting typological features in WALS (Bjerva and Augenstein, 2018a). As their setup is different, we use their highest reported score as a baseline.

Feature Prediction Results. Table 1 contains the results from feature prediction across the chapters outlined in WALS. Our implementation is able to predict features across categories above baseline levels. At increasing numbers of implicants, prediction power tends to increase. This is not the case for all feature categories, however. One such case is Nominal Syntax, in which performance peaks at 3 implicants. This is expected, as correlations only exist between some features, thus at a certain point access to more typological features no longer helps performance. Note that although the baseline numbers are based on predicting the same features as our model, the baseline models do not observe the same features during prediction - for instance Baseline #4 does not make predictions based on other feature values, but is trained on one feature at a time.

4.2 Discovering Typological Implications

Having established that our method bests several competitive baselines for prediction of typological features, we next look at what implications our probabilisation of typology allows us to find. We search for those conditional probabilities where the quantity $|p(\pi_i \mid \pi_j) - p(\pi_i)|$ is statistically significantly greater than 0, as found with an independent

---

[1] We use the original implementation of PRA available here: https://github.com/noon99jaki/pra

#	Implicant ⊃ Implicand
1*	Postpositions ⊃ Genitive-Noun (Greenberg #2a)
2*	Postpositions ⊃ OV (Greenberg #4)
3	OV ⊃ SV
4*	Postpositions ⊃ SV
5*	Prepositions ⊃ VO (Greenberg #4)
6*	Prepositions ⊃ Initial subord. word (Lehmann)
7*	Adjective-Noun, Postpositions ⊃ Demonstrative-Noun
8*	Genitive-Noun, Adjective-Noun ⊃ OV
9	SV OV Noun-Adjective ⊃ SOV
10	Degree word-Adjective VO and Noun–Relative Clause SVO ⊃ Numeral-Noun
11	SOV OV and Relative Clause–Noun Adjective-Degree word ⊃ Noun-Numeral

Table 2: Hand-picked implications. In cases where the same is covered by Daumé III and Campbell (2007), we borrow their analysis (marked with *).

two-tailed t-test.[2] After adjusting for multiple tests with the Bonferroni correction, we report those implications where $p < 0.05$. We report the full list of implications found by our model in the Supplements and show a subset of these in Table 2.[3] We note that we are able to find the same implications listed by Daumé III and Campbell (2007), some of which are listed in the table. These implications include Greenberg universals (Greenberg, 1963), showing that our approach to probabilisation of linguistic universals is suitable to replicate previous work.

Transitivity across implications At first glance, it is not clear why postpositions should imply SV word order, as stated in #4. Yet, #2 is a well-established universal (Greenberg, 1963) and #3 comes with strong statistical evidence: SV order is much more frequent than VS word order in OV languages (98.44% of these are predominantly SV). Our model has thus used transitive reasoning of the form *if $A \supset B \wedge B \supset C$ then $A \supset C$* to find #2.

The power of multiple implicants Implications #10 and #11 concern the order between nouns and their numeral modifiers. The two main alternatives here, Noun-Numeral and Numeral-Noun are of comparable frequency in WALS; they occur in 607 and 479 languages, respectively, i.e. Noun-Numeral holds the majority with only 55%. If we consider each of the three implicants listed in implication #11 on their own, the strongest statistical

power goes to the *Degree word–Adjective* feature: conditioned on this feature, the Numeral-Noun order holds in 79% of the relevant languages. The combination of all three implicants, on the other hand, results in a subset of languages with 91% Numeral-Noun order. The Numeral-Noun order can thus be implied with considerably more confidence from a combination of multiple implicants.

5 Related Work

Typological implications outline the space of possible languages, based on evidence from observed languages, as recorded and classified by linguists (Greenberg, 1963; Lehmann, 1978; Hawkins, 1983). While work in this direction has been manual, typological knowledge bases do exist now (Dryer and Haspelmath, 2013; Partick Littel and Levin, 2016), which allows for automated discovery of implications. Although previous computational work exists (Daumé III and Campbell, 2007), we are the first to introduce a probabilisation of typological implications.

In addition to work on finding implications based on known features, there is an increasing amount of work on computational methods to discovering typological features (Ponti et al., 2018). Work in this area includes unsupervised discovery of word order (Östling, 2015) or other linguistic features (Asgari and Schütze, 2017), typological probing of language representations (Bjerva et al., 2019b; Beinborn and Choenni, 2019), and several papers attempt to predict typological features in WALS (Georgi et al., 2010; Malaviya et al., 2017; Bjerva and Augenstein, 2018a,b; Cotterell and Eisner, 2017, 2018; Bjerva et al., 2019a).

6 Conclusions

We defined the notion of probabilistic implications, and presented a computational model which successfully identifies known universals, including Greenberg universals, but also uncovers new ones, worthy of further investigation by typologists. Additionally, our approach outperforms strong baselines for prediction of typological features.

---

[2] Future work will make use of a non-parametric test, whose details we are still working out.

[3] Also on `bjerva.github.io/imp_acl19.pdf`.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Lisa Beinborn and Rochelle Choenni. 2019. Semantic Drift in Multilingual Representations. *arXiv preprint arXiv:1904.10820*.

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 Computatinal Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, 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, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019b. What do Language Representations Really Represent? *Computational Linguistics*, 45(2):381–389.

Noam Chomsky. 1981. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic Typology: Deep Generative Models of Vowel Inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2018. A Deep Generative Model of Vowel Formant Typology. 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 37–46, New Orleans, Louisiana. Association for Computational Linguistics.

William Croft. 2002. *Typology and Universals*. Cambridge University Press.

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.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing Language Similarity across Genetic and Typologically-Based Groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 385–393, Beijing, China. Coling 2010 Organizing Committee.

Joseph Harold Greenberg. 1963. *Universals of Language*. MIT Press.

John A. Hawkins. 1983. *Word Order Universals: Quantitative analyses of linguistic structure*. Academic Press.

Ni Lao and William W. Cohen. 2010. Relational Retrieval Using a Combination of Path-constrained Random Walks. *Machine Learning*, 81(1):53–67.

Ni Lao, Tom Mitchell, and William W. Cohen. 2011. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Winfred P. Lehmann. 1978. Syntactic Typology. *Studies in the Phenomenology of Language*, pages 3–55.

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, Brussels, Belgium. Association for Computational Linguistics.

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.

Richard E. Neapolitan. 2004. *Learning Bayesian Networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ.

Robert Östling. 2015. Word Order Typology through Multilingual Word Alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *ACL*, pages 644–649.

David Mortensen Partick Littel and Lori Levin, editors. 2016. *URIEL Typological Database*. Carnegie Mellon University, Pittsburgh.

Judea Pearl. 1982. *Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach.* Cognitive Systems Laboratory, School of Engineering and Applied Science.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *arXiv preprint arXiv:1807.00914*.

Dingquan Wang and Jason Eisner. 2016. The Galactic Dependencies Treebanks: Getting More Data by Synthesizing New Languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.