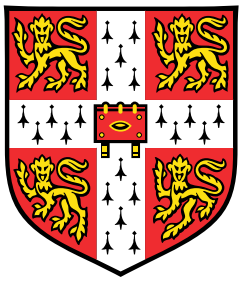
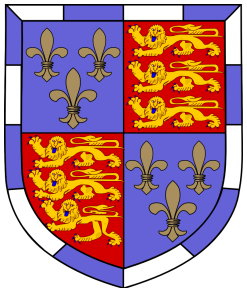


Inductive Bias and Modular Design for Sample-Efficient Neural Language Learning



Edoardo Maria Ponti

Supervisor: **Prof. Anna Korhonen**

Co-supervisor: **Dr. Ivan Vulić**

Theoretical and Applied Linguistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St John's College July 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Edoardo Maria Ponti
July 2020

Inductive Bias and Modular Design for Sample-Efficient Neural Language Learning

Edoardo Maria Ponti

Abstract

Most of the world's languages suffer from the paucity of annotated data. This curbs the effectiveness of supervised learning, the most widespread approach to modelling language. Instead, an alternative paradigm could take inspiration from the propensity of children to acquire language from limited stimuli, in order to enable machines to learn *any* new language from few examples. The abstract mechanisms underpinning this ability include 1) a set of in-born inductive biases and 2) the deep entrenchment of language in other perceptual and cognitive faculties, combined with the ability to transfer and recombine knowledge across these domains. The main contribution of my thesis is giving concrete form to both these intuitions.

Firstly, I argue that endowing a neural network with the correct inductive biases is equivalent to constructing a prior distribution over its weights and its architecture (including connectivity patterns and non-linear activations). This prior is inferred by 'reverse-engineering' a representative set of observed languages and harnessing typological features documented by linguists. Thus, I provide a unified framework for cross-lingual transfer and architecture search by recasting them as hierarchical Bayesian neural models.

Secondly, the skills relevant for different language varieties and different tasks in natural language processing are deeply intertwined. Hence, the neural weights modelling the data for each of their combinations can be imagined as lying in a structured space. I introduce a Bayesian generative model of this space, which is factorised into latent variables representing each language and each task. By virtue of this modular design, predictions can generalise to unseen combinations by extrapolating from the data of observed combinations.

The proposed models are empirically validated on a spectrum of language-related tasks (character-level language modelling, part-of-speech tagging, named entity recognition, and common-sense reasoning) and a typologically diverse sample of about a hundred languages. Compared to a series of competitive baselines, they achieve better performances in new languages in zero-shot and few-shot learning settings. In general, they hold promise to extend state-of-the-art language technology to under-resourced languages by means of sample efficiency and robustness to the cross-lingual variation.

Acknowledgements

I would like to thank Professor Anna Korhonen for giving me the incredible opportunity to pursue a PhD in her group and encouraging my passion for research. Her guidance as a supervisor has been invaluable. Moreover, I am indebted to Ivan for his patience in listening to my weekly ravings and for his friendship full of Tolkienian riddles: *yéni ve lintë yuldar avánier*!

I dedicate this thesis to Olga. Her sweetness, dedication, curiosity, and wit filled a void in me that I never hoped to fill before meeting her. I am looking forward to our next adventures together around the world!

I thank my mother Cinzia, my father Gigi, and my sister Chiara. Their unwavering love has given me strength when in hardship and refuge when in doubt.

Especially important in my development as a researcher were Ryan, who saved me from becoming an armchair philosopher navel-gazing my life away in a lavish Oxbridge study, and Roi, whose passion in voicing praise and criticism I really admire.

My sincere gratitude goes to my fellow students at the Language Technology Lab: Daniela, Gamal, Billy, and Milan for welcoming me with open arms, and Sangseo, Costanza, Yi, Flora, Victor, Marinela, and Fangyu for all the laughs and serious discussion.

I also wish to thank the other researchers I had a chance to collaborate with during my graduate studies: Katia Shutova, Goran Glavaš, Anne Lauscher, Aishwarya Kamath, Jonas Pfeiffer, Dieuwke Hupkes, Elia Bruni, and Thierry Poibeau. It is hard to overstate how much I owe to their insights and help.

I should at least mention the committees of the Samuel Butler Room Society at St. John's College and of the Cambridge University Italian Society as an inexhaustible source of joy and entertainment during my time in Cambridge. Special thanks to Fulvio, Dario, Concetta, Francesca, Harriet, Victor, and Iacopo. *Ad maiora* Ludovico and Chiara!

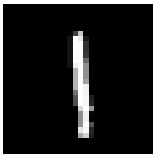
Also, I wish to thank the Apple group in Cupertino, and in particular Sid, Ravi, and Michael, for an unforgettable Californian experience.

Finally, I thank the *habitués* of my acknowledgements, the friends who accompanied me throughout my life and whose bond is renewed at each encounter.

Contents

List of Figures

List of Tables

Chapter 

Introduction

1.1 Motivation

Current machine learning models rely on abundant data and often assume that training and evaluation data are *i.i.d.* Consequently, they struggle in natural language processing tasks, as most languages suffer from a dearth of training examples and the typological variation makes each language differently distributed (Linzen et al., 2016; Yogatama et al., 2019). In this thesis, I argue that these difficulties can be overcome by taking inspiration from how humans learn and use language. In Section 1.1.1, I first draw a comparison with language acquisition in children, which is characterised by limited stimuli and the entrenchment of the language faculty inside the human perceptual, cognitive, and communicative system. This contrasts with machine learning models, which tend to be learned from scratch and in isolation. Afterwards, in Section 1.1.2 I argue that the above-mentioned limitations arise from this misalignment; however, they can be mitigated by constructing an adequate inductive bias and a mechanism to disentangle and recombine knowledge. To do so, in Section 1.1.3 I propose a unified Bayesian neural framework that attains several desirable properties, including sample efficiency, resilience to catastrophic forgetting, compositional generalisation to new domains, and robustness to uncertainty. Finally, in Section 1.1.4 I show how the proposed approach not only facilitates natural language processing applications, but has also the potential to shed light on several scientific challenges of linguistic typology.

1.1.1 Two Alternative Paradigms

The languages spoken and signed around the world vary remarkably in their structures and lexicon. Yet, human children can master any of them swiftly based on a limited amount of stimuli (Chomsky, 1980). Such 'language instinct' cannot be hard-wired into the human genome, as this would not account for the learners' *flexibility* in the face of synchronic variation and diachronic change. In fact, the notion of innate inviolable constraints on possible language structures has been disavowed (Perfors et al., 2011) because of the exceptions to any proposed universal rule (Evans and Levinson, 2009).

On the other hand, it would be hard to reconcile the sample *efficiency* of learning with a purely empirical process. Hence, one must posit an inborn inductive bias that expedites language acquisition from experience (Zador, 2019). This idea is corroborated by the fact that cross-lingual variation does not behave randomly, but rather follows precise universal tendencies (Comrie, 1989). These are partly attributed to the embodied nature of language (Majid et al., 2007). For example, the neuro-physiology of vision constrains possible patterns in the lexical field of colours (Kay and McDaniel, 1978) as much as the shape of the oral cavity determines plausible inventories of vowels in phonology (Lindblom, 1986). Hence, language is uniquely *intertwined* with systems of perception, cognition, and communication that humans evolved. Children draw on these to acquire and use a language.

Unfortunately, models of language based on *machine learning* are hardly as flexible, efficient, and well-integrated. Admittedly, artificial neural networks achieve state-of-the-art (and sometimes super-human) performance on most benchmarks for natural language processing (LeCun et al., 2015; Wang et al., 2019). Nevertheless, this success is limited to a handful of tasks and language varieties, since it is predicated on a series of conditions that are generally impossible to meet. In particular, deep learning models require massive amounts of labelled data for supervised training. Since their parameters are usually initialised randomly, the burden of learning falls entirely on experience. In practice, this approach raises insurmountable difficulties: most of the world's languages lack labelled data, whose creation is expensive, and often even digital texts (Kornai, 2013). As a result, sample-efficient learning from limited evidence, known as zero-shot and few-shot learning, is paramount to enhancing the outreach of machine learning to under-resourced languages.

Besides, models are usually dedicated to the solution of an individual task, which demands only a fraction of the skills necessary to engage in language as a whole. Often, this is equivalent to assuming a stationary data distribution between training and evaluation. According to Linzen (2020), this fosters the selection of neural architectures that are low-

bias and become over-sensitive to the idiosyncratic features of a particular dataset. When a shift occurs in the distribution of evaluation data, reliance on training data artefacts can lead to dramatic drops in performance (Niven and Kao, 2019). Instead, general linguistic intelligence should allow for generalising in unseen situations by abstracting from previously gained knowledge (Yogatama et al., 2019), while preventing 'catastrophic forgetting' when adapting to novel evidence (French, 1999). In other words, if a neural model is optimised for a sequence of tasks, it should not only retain previous information, but also exploit it to have a head-start on novel tasks.

An attractive solution to the limitations of data paucity and task fragmentation is *knowledge transfer* across language varieties, tasks, and modalities (such as vision, speech, and motion) (Caruana, 1997; Ruder et al., 2019b), which reflects the synergy of different linguistic skills and takes advantage of similarities among language varieties. Contextualised word embeddings are a recent successful example of cross-task transfer (Devlin et al., 2019; Peters et al., 2018; Raffel et al., 2019). In particular, these representations are pre-trained on unlabelled texts through language modelling and subsequently fine-tuned on annotated data from supervised learning tasks. Cross-lingual transfer instead leverages data from resource-rich languages to perform inference in resource-lean languages through annotation projection, multilingual representation learning, or translation (Conneau et al., 2018; Hwa et al., 2005; Tiedemann, 2015; Yarowsky et al., 2001; Zeman and Resnik, 2008, *inter alia*). Transfer can be carried out simultaneously across both languages and tasks for zero-shot predictions on languages that have raw texts available but no labelled examples (Conneau and Lample, 2019; Pires et al., 2019).

Knowledge transfer alone, however, often yields unsatisfactory results, as it does not guarantee sample efficiency nor forestall generalisation errors. In fact, effective fine-tuning still hinges upon the availability of a significant amount of in-domain examples (Ravi and Larochelle, 2017; Vinyals et al., 2016), and state-of-the-art pre-training encoders are especially prone to over-fit to training data (Niven and Kao, 2019).

1.1.2 Sample Efficiency and Generalisation

In this thesis, I argue that these challenges in multilingual natural language processing can be addressed by taking inspiration from traits of language acquisition in humans, thus narrowing the chasm between these two paradigms. In particular, is it possible to individuate a bulk of universal linguistic knowledge that can be re-elaborated and preserved for a model to become competent in new languages quickly? This corresponds exactly to the idea of imbuing models with the correct inductive bias; rather than evolving

it through natural selection, however, this artificial counterpart should be distilled by 'reverse-engineering' data from other language varieties.

To the acquainted reader, this idea is certainly reminiscent of the intuitions behind meta-learning and continuous learning within the scope of cross-task transfer: neural *meta-learning*, or 'learning to learn' (Finn et al., 2017), aims at achieving sample efficiency by finding model parameter values optimised for generalisation, i.e. for which local optima of any new task are found just a few steps of gradient descent away. The equal and symmetric problem of remembering old knowledge while incorporating original one instead is tackled by *continuous learning* through elastic weight consolidation (Kirkpatrick et al., 2017), model compression (Schwarz et al., 2018), or memory blocks (Grave et al., 2017). In addition to accommodating this set of ideas to the peculiarities of cross-lingual transfer, in a realistic setting the coarse-grained features of the target language to be learned should also blend into the inductive bias for language as side information. In fact, we are never completely in the dark regarding such features, as they are often documented by linguists in typological resources based on the comparison of the world's languages (Dryer and Haspelmath, 2013; Littell et al., 2017).

Moreover, the parallel established with the innate component of natural languages helps to take the formulation of inductive bias a step further. This component is informed by our genome, which steers the brain connectivity patterns rather than connectivity strengths (Zador, 2019). From a modelling perspective, focusing exclusively on weight initialisation is only part of the story. In fact, each deep neural network defines a class of non-linear functions (MacKay, 2003, ch. 45). Making assumptions about a model architecture (activation functions, layer size and depth) through fixed hyper-parameters narrows down the choices for learnable functions, possibly excluding those most suitable for fast adaptation. While a toolbox of techniques known as *neural architecture search* allows for gradient-based joint inference over weights and architectures (Liu et al., 2019; Xie et al., 2019), it remains unclear how to deploy them onto a few-shot learning setting, in order to equip a model with an inductive bias encompassing the model architecture.

Finally, to achieve satisfactory generalisation capabilities, an ideal model should mirror the integrated but modular system in which language is deep-seated. In fact, neural networks benefit from sharing the same architecture across tasks, language varieties, and modalities, which enables weight sharing (McCann et al., 2018; Raffel et al., 2019). Reserving a set of 'private' weights for each specific task–language–modality combination does not allow for borrowing strength from the others and leads to a proliferation of the number of parameters. On the other hand, it is unreasonable to lump together all linguistic knowledge into a single, monolithic set of parameters. Instead, only the

relevant knowledge for a specific combination should be ideally accessed, generating parameters 'on-demand'. For instance, given training data for named entity recognition (NER) in Vietnamese and for part-of-speech (POS) tagging in Wolof, a model should perform accurate predictions for NER in Wolof. Similarly, when transferring knowledge from textual question answering to visual question answering, a model should retain the information associated with the task variable, while dispensing with the information about the current modality.

Overall, the **desiderata** for neural models of language can be summarised as follows:

D1: *Sample efficiency* to learn from few examples by virtue of an inductive bias, especially in light of the data paucity in most of the world's languages;
D2: Avoiding *catastrophic forgetting* in order to preserve acquired knowledge and enable continuous learning.
D3: Ability to *disentangle and recombine* knowledge from past experience when facing unprecedented combinations of languages, tasks, and modalities;
D4: *Robustness* to data distribution shifts between training and evaluation, 'failing loudly' whenever uncertainty prevents any reasonable prediction.

1.1.3 A Unified Bayesian Neural Framework

This thesis adopts a Bayesian perspective towards modelling and inference in neural networks (Blundell et al., 2015; Kingma and Welling, 2014), as this satisfies all the above-listed desiderata inside a unified framework. Ingrained in it, in fact, is the requirement to explicate priors, which can be taken to represent inductive biases. In this thesis, I argue for inferring a posterior over weights and model architectures through Laplace approximations (MacKay, 1992) or variational approximations (Wainwright and Jordan, 2008) from observed language. This distribution can subsequently serve as a prior for maximum-a-posteriori inference or model averaging when the model is exposed to few examples of a held-out language.

Results from the following chapters in character-level language modelling on a sample of 77 languages demonstrate the superiority of an expressive prior on neural weights over uninformative priors and unnormalisable priors (i.e., the widespread 'fine-tuning' approach) in both zero-shot and few-shot settings. A suitable prior is not only superior to learning from scratch in terms of performance and sample efficiency, but also prevents catastrophic forgetting compared to maximum likelihood estimates by virtue of Bayesian updating. The constructed prior will be shown to capture both universal and language-

specific phonotactic knowledge for modelling character sequences. Similarly, an informed prior over neural architectures and weights imbued with world knowledge improves common-sense reasoning on a sample of 12 languages.

A second advantage of the Bayesian approach is explicitly controlling for the variables at play and their pairwise (in)dependence in the form of a graph. In this thesis, I maintain that the space of neural parameters is inherently structured as a tensor where each cell is a possible combination of tasks, languages, and modalities. Each dimension instead represents an autonomous aspect of linguistic knowledge. Hence, I propose a generative Bayesian model of such a space that factorises into distinct latent variables for each task and language. Since some of their combinations are observed, the knowledge relevant to each of them can be distilled and stored. Subsequently, the aspects relevant to unobserved combinations can be accessed and recombined appropriately. Results over a range of tasks (such as part-of-speech tagging and named entity recognition) demonstrate that this supplies a mechanism to achieve better generalisation than established methods for cross-lingual and cross-task transfer.

Thirdly, Bayesian models yield smoother predictive distributions, which better reflect the model uncertainty during prediction. Indeed, a notable limitation of point estimate methods is their tendency to assign most of the probability mass to a single class even in scenarios with high uncertainty. Zero-shot transfer is one such scenario, especially when it involves drastic distribution shifts in the data (Rabanser et al., 2019). The ability to 'fail loudly' in such cases makes the prediction more robust. In this thesis, I take advantage of one of the most prominent features of Bayesian inference, namely model averaging, to show how low entropy in the (approximated) predictive distributions correlates almost perfectly with high performance. This introduces the possibility to refrain from making predictions in domains where the model confidence is insufficient.

Overall, this unified neural Bayesian framework provides the following **solutions** to the desiderata individuated in Section 1.1.2:

S1: Constructing a *prior distribution* over neural parameter weights and architectures from observed languages and *side information* from typological features ameliorates sample efficiency in learning a new language from few examples.
S2: If the prior acts as a *regulariser*, it averts the possibility that the information about observed languages is overridden via catastrophic forgetting.
S3: A mechanism enabling generalisation is brought about via a modular design, by *factorising* the neural parameter space into variables accounting for specific aspects of linguistic knowledge, whose dependencies are articulated via a graph.
S4: Uncertainty in a domain is mirrored by the entropy of the predictive distributions obtained through model averaging. This makes predictions more robust than maximum likelihood estimates.

1.1.4 Data-driven Linguistic Typology

In addition to improving sample-efficient language learning in neural networks, the proposed approach ushers in new possibilities for theoretical linguistics. In particular, it allows for simulating aspects of language acquisition and cross-lingual variation in a data-driven fashion. Traditional studies classify languages into 'types', a sort of language taxonomy. Principles of human cognition (such as economy and iconicity) are sought through the analysis of the cross-lingual patterns of these types (Croft, 2002). However, types are often coarse-grained and partly arbitrary, possibly distorting subsequent analyses. Moreover, they need to be manually documented, which creates a bottleneck reducing the coverage of language samples. Instead, the process of typological analysis could be automatised and grounded empirically without an intermediate taxonomic level, by inferring the correct set of inductive biases that explains the cross-lingual variation directly from textual data (or even non-linguistic data such as vision and communication between artificial agents). Hence, probing the constructed priors holds promise to unveil cognitive dynamics in language learning.

1.2 Thesis Outline

This thesis is organised into 6 chapters. After the introduction in Chapter 1 and background in Chapter 2, the question of constructing neural networks inductively biased towards language is contemplated for weights in Chapter 3 and for architectures in Chapter 4. Chapter 5 concerns the implementation of neural networks with a modular

design through parameter factorisation. Finally, I draw some conclusions about the success of these undertakings in Chapter 6. This thesis contains the following original contributions, ordered by chapter:

2 I overview the connection between language variation and acquisition, to substantiate the claim that both stem from principles of human perception, cognition, and communication. Afterwards, I argue that the same principles should inspire modelling and inference in natural language processing. To this aim, I lay the foundations for a unified framework of knowledge transfer, neural architecture search, and parameter factorisation through Bayesian neural networks.

3 This chapter focuses on character-level, open-vocabulary language modelling, benchmarking models on the largest sample of typologically diverse languages (77) to date. In particular, I adapt elastic weight consolidation (Kirkpatrick et al., 2017) to cross-lingual transfer for the first time, outperforming the established method of 'fine-tuning' in zero-shot and few-shot learning settings. Moreover, I devise several methods to condition parameters on typological features to integrate linguistic side information seamlessly. In particular, I assess the viability of feature concatenation and hyper-networks for parameter generation. Finally, I probe the learned posterior over weights to show that it is imbued with universal phonotactic knowledge.

4 In this chapter, I explore the idea of inductively biased learning via a prior over architectures. To do so, I recast differentiable neural architecture search (NAS) as hierarchical Bayes, whereby weights are generated based on samples from a categorical distribution over layer connections and non-linear activations. Moreover, by virtue of this interpretation, I show how the architecture prior facilitates zero-shot and few-shot learning in a novel challenging benchmark. In particular, I create Cross-lingual Choice of Plausible Alternatives (XCOPA), a typologically diverse multilingual dataset for causal commonsense reasoning in 11 languages.

5 In this chapter, I consider the space of neural parameters as inherently structured, as it results from the possible combinations of specific tasks (POS tagging and NER) and languages (33). Hence, I propose a Bayesian generative model where such space is factorised into latent variables for each language and each task. By performing variational inference from data in observed combinations, I report gains in zero-shot sequence classification on held-out combinations at prediction time. Finally, I show how the entropy of the (approximate) predictive distributions anti-correlates with performance metrics.

6 I review the main experimental results, and examine to which extent they corroborate the core idea of this thesis, namely that inductive bias and modular design draw machine learning models closer to the above-mentioned desiderata of sample efficiency, resilience to catastrophic forgetting, generalisation, and robustness to uncertainty. Finally, I conjecture about future work, for instance: i) to construct priors from artificial languages emerging from multi-agent communication rather than other natural languages; ii) to broaden parameter factorisation to multiple modalities (such as vision and speech) as well as to other neural components (such as Adapter layers).

1.3 Publications

I ran all the experiments reported in this thesis, with the following exceptions. Daniela Gerz evaluated n-gram and neural language models on a sample of 40 languages, whereas I carried out the analysis on top of these results. Goran Glavaš implemented the baseline models for the XCOPA dataset. I am indebted to all co-authors for their help. This thesis features content from the following papers (ordered by chapter of appearance):

- **Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing.** Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. *Computational Linguistics* 45(3):559–601.
  ✍ Chapter 2

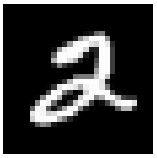
- **Isomorphic Transfer of Syntactic Structures in Cross-lingual NLP.** Ponti, Edoardo Maria, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1531–1542.
  ✍ Chapter 2

- **Towards Zero-shot Language Modeling.** Ponti, Edoardo Maria, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2893–2903.

✍ Chapter 3

- **XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning.** Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362-2376.
  ✍ Chapter 4

- **Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages**. Ponti, Edoardo Maria, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. *Transactions of the Association for Computational Linguistics*.
  ✍ Chapter 5

Chapter 

Background

2.1 Language Variation and Acquisition

The world is blessed with a wealth of languages, although it is far from straightforward to estimate how many. In fact, drawing boundaries between language varieties is based on their mutual intelligibility, which is gradient rather than clear-cut. As a result, the total number of languages falls somewhere between 7117 (Lewis et al., 2016) and 7604 (Hammarström et al., 2020).[1] At first sight, the differences across languages are striking; however, upon closer inspection, deep connections can be unveiled. For instance, while wandering around the linguistically diverse region of the Caucasus, one may overhear sentence 2.1 in Lezgian (Haspelmath, 1993, p. 148), or sentence 2.2 in Georgian (Boeder, 2000, pp. 285–286):[2]

(2.1) *Qe* *sobranie* *že-da-**lda**.*
today meeting be-FUT-QUOT
'Apparently, there will be a meeting today.'

(2.2) *Tovl-i* ***mosula**.*
snow-NOM come.PERF
'It must have snowed.'

Although the *forms* of the linguistic units in these examples are entirely different because of the arbitrariness of the lexical sign (Saussure, 1916), some of them fulfil the same

[1] These counts include only languages traditionally spoken by a community as their principal means of communication, and exclude unattested, pidgin, whistled, and sign languages.

[2] All examples are glossed according to the Leipzig rules (Comrie et al., 2008).

function. Namely, both the quotative suffix *-lda* on the main verb in sentence 2.1 and the perfect tense form of the verb *mosula* in sentence 2.2 convey evidentiality. Evidentiality indicates the source of information for a statement: in this case, it is indirect, such as hearsay or circumstantial inference by the speaker (de Haan, 2013). By broadening our sample of languages, we may notice that other languages express indirect evidentiality, such as Kannada in sentence 2.3 (Sridhar, 1990, p. 3) and Dutch in sentence 2.4:

(2.3) *Nimma pustaka avara hattira illav-**ante**.*
your book he.POSS near NEG-QUOT
'Allegedly, your book is not with him.'

(2.4) *Het **moet** een goede film zijn.*
it MOD a good film be.INF
'It must be a good film.'

Although the form of *-ante* in Kannada is unrelated to *-lda* in Lezgian, they are equivalent because they are both affixes. Otherwise stated, languages may adopt the same formal *strategy* (although not necessarily the same lexical unit) to codify a specific function (Croft et al., 2017). On the other hand, Dutch resorts to a previously unexpected and exotic strategy, the modal verb *moet*.[3] Obviously, classifying languages into 'types' according to their most frequent strategy requires a set of cross-lingually valid categories: in this case, affix, inflectional tense, and modal verb. Concretely, there is no finite set of such categories that can be fixed in advance (Haspelmath, 2007); rather, they are progressively refined as new evidence becomes available (Bickel, 2007, p. 248).

2.1.1 Typological Universals

Performing this kind of comparison systematically across functions based on a representative sample of the world's languages is the goal of linguistic typology (Comrie, 1989; Croft, 2002). The type of each language is documented in typological databases, such as the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013).[4] For instance, at least 6 types are attested for the function of evidentiality according to WALS: in Figure 2.1, each language on the world map is colour-keyed based on its strategy.[5] As it emerges, languages spoken in the same area may tend to share the same strategy.

[3]Exotic indeed: this strategy is common only to 7 languages in a sample of 418, and they are mostly concentrated in Western Europe

[4]For a comprehensive list of typological databases, consult Ponti et al. (2019a).

[5]The language locations in the plot span Voronoi cells rather than dots as proposed by McNew et al. (2018) to visualise areal contiguity.

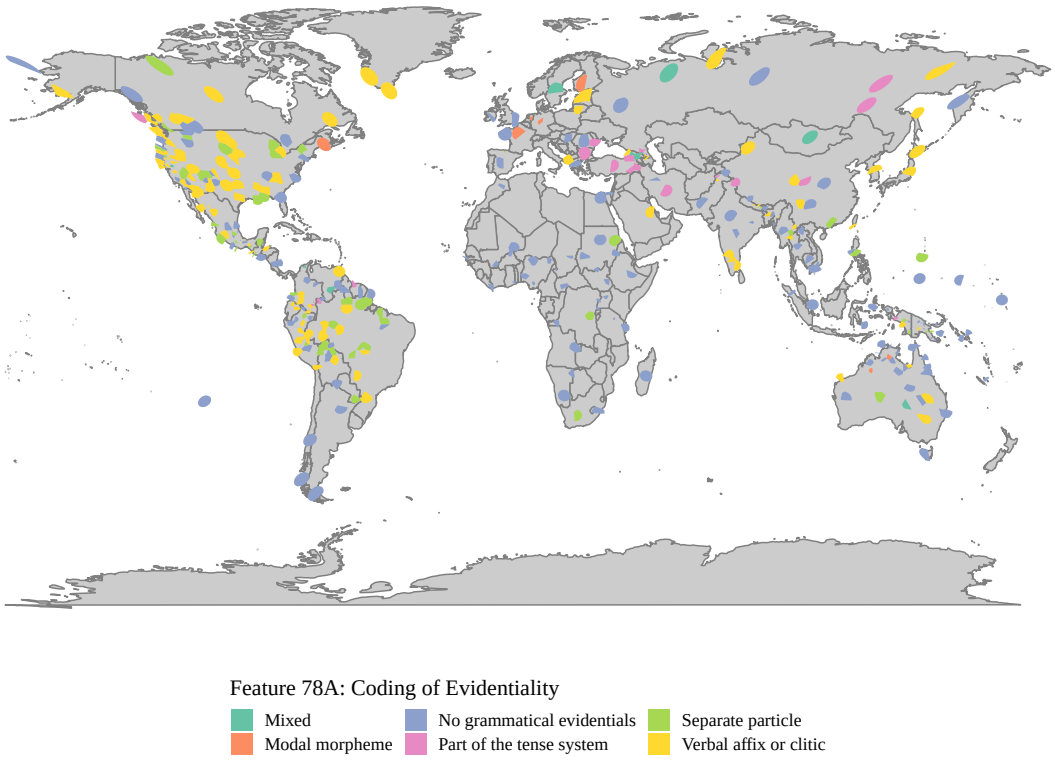


Figure 2.1 Map of the types of strategies to code evidentiality in the world's languages according to WALS (Dryer and Haspelmath, 2013).

Moreover, typological databases account for the variation in other levels of linguistic description. The *structural* level is concerned purely with the form of linguistic units (such as phonemes, morphemes, words, clauses). For instance, languages can be classified based on whether grammatical morphemes tend to be isolated, concatenated, or fused with their root (Sapir, 2014 [1921], p. 128). The *semantic* level, instead, focuses on the allocation of concepts into categories in the lexicon (Evans, 2011). For instance, languages can be classified in terms of the granularity of a semantic field: they can either distinguish FINGER from ARM through separate words, or cover both under a single umbrella term. Henceforth, I refer to any aspect regarding which languages can be compared as a *typological feature*.

As a result, a typological database can be conceived as a binary matrix where each row is a language $\ell \in L$, and each column is a (binarised) feature $f \in F$ (Georgi et al., 2010). Each cell contains a 1 if a language belongs to the corresponding type, and a 0 otherwise. Hence, a language can be represented as a vector of typological features

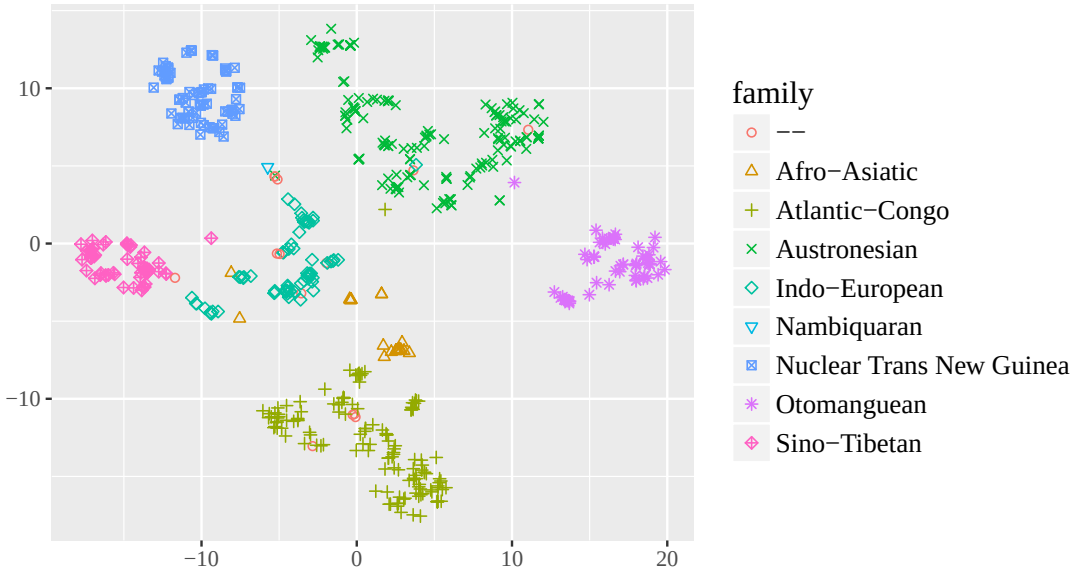


Figure 2.2 Language features from WALS dimensionality-reduced with t-SNE.

$\mathbf{t}_\ell \in [0,1]^{|F|}$. The feature vector of each language in WALS is visualised in Figure 2.2 after being reduced to 2 dimensions through t-SNE (Maaten and Hinton, 2008).[6] Note, incidentally, that points aggregate in space according to family membership.

The information contained in databases, however, is far from exhaustive. First, it is sparse and skewed, as some languages and some features are better documented than others. As a result, many cells in the matrix remain blank. Second, most typological databases are coarse-grained as they fail to account for feature variations *within* each language: reporting only the majority values overshadows the fact that multiple strategies are often attested simultaneously, although with different frequencies. Further challenges are posed by restricted feature applicability and feature hierarchies, which introduce redundancies and nonassignable entries (Ponti et al., 2019a).

Despite all these limitations, the copious evidence provided by typological databases can be examined to ascertain general patterns of cross-lingual variation. As I already mentioned above, the distribution of features across languages is not random, but rather depends on family and area. In fact, similar strategies can be inherited from a common

[6] Missing values are populated automatically through a weighted nearest neighbour algorithm (Littell et al., 2017).

ancestor (Ross, 1997), or borrowed by contact with a neighbour. For instance, in all languages part of the Eastern Tucanoan family, spread across the western Amazon forest, evidentiality is part of the tense system. As an example of geographic percolation, most languages in Africa do not grammaticalise evidentiality at all, as shown in Figure 2.1. Moreover, since family and area explain part of the variation of other features, too, they influence the entire 'profile' of a language, which results in similar feature vectors (see Figure 2.2 again).

What is more, typological features are *not* independent from each other given family and area. Indeed, even accounting for these variables (Bakker, 2010), typological features turn out to display a high degree of 'solidarity': the presence of one feature may implicate another (in one direction or both). The discovery of these patterns, called *universals*, is owed to Greenberg (1966). For example, if adpositions precede their noun, then genitive-like modifiers tend to follow their noun, and vice versa. It is worth stressing that these implications are not deterministic (Corbett, 2010), as exceptions are known for most of the universals if understood as absolute (Evans and Levinson, 2009).

The dependency of each feature on both area / family and other features is due to the hybrid nature of language, which involves both cultural and biological components (Durham, 1991). Cross-lingual variation can therefore be explained from two complementary perspectives: on the one hand, event-based theories focus on the *diffusion* of features due to family inheritance or areal percolation, accounting for their propagation or extinction (Bickel, 2015). On the other hand, functional theories emphasise the influence of cognitive and communicative principles in the *origins* of innovations among the features of a language (Croft, 1995, 2000).

Ultimately, the origin of innovations traces back to the individual speaker. This leads to the generation of multiple strategies within a language community. The selection among these variants is socially governed, for instance according to the prestige of the speakers adopting it (Herzog et al., 1968). Since this selection is completely independent of the innovation, however, language-internal variation is reflected faithfully in cross-lingual variation (Croft, 2001, p. 107). In fact, typological universals can be considered as recurrent solutions in time and space by individuals, and outliers as rare happenstances triggered by unlikely preconditions (Evans and Levinson, 2009).

Hence, typological universals can shed light on the shared principles underlying the grammatical knowledge of individuals. Communicative principles favour linguistic expressions that are frequently used or easy to process (Cristofaro and Ramat, 1999; Haspelmath, 1999). At the same time, cognitive principles constrain the mapping between semantic functions and formal strategies. In particular, functions can be arranged into a

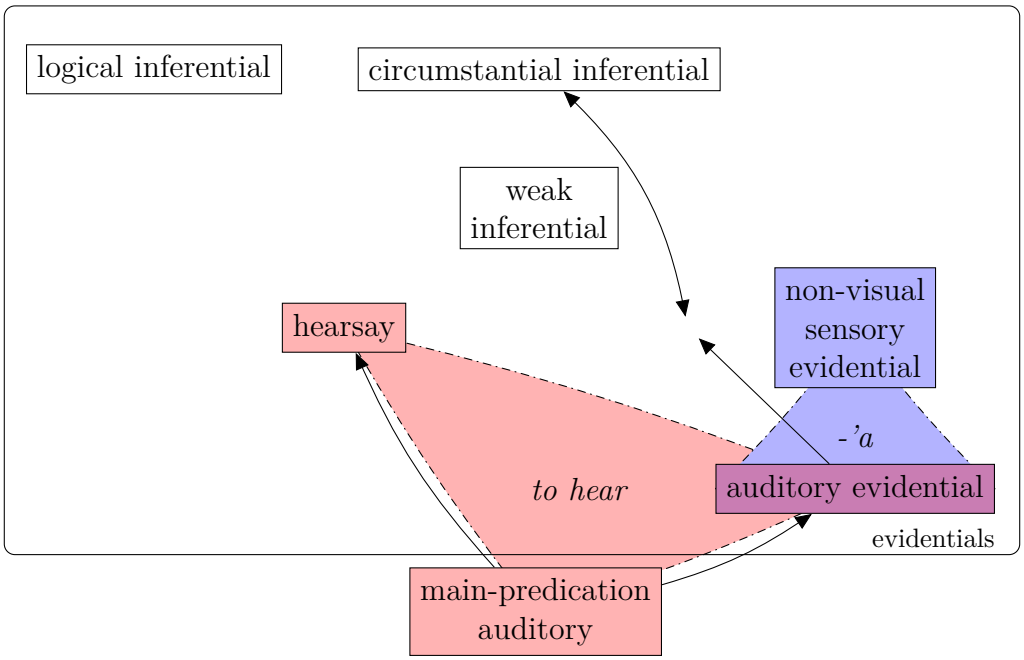


Figure 2.3 Simplified semantic map of evidentials according to Anderson (1986). Boxes are functions, arrows common diachronic trajectories of semantic shift.

'semantic map' where each language-specific form can express only a range of language-independent functions lying in a contiguous region (Haspelmath, 2003). Discrete functions (points in the meaning space) are distinguished if and only if there exist a pair of languages whose forms draw different boundaries with respect to them. For instance, consider the (simplified) semantic map of evidentials according to Anderson (1986) in Figure 2.3. Possible functions are surrounded by boxes; those of concern here can be exemplified as follows:

(2.5) MAIN PREDICATION: *The prophet* ***heard*** *a divine voice.*

HEARSAY: *The journalist* ***heard*** *that Haruki Murakami would win the Nobel prize.*

AUDITORY EVIDENTIAL: *The rebel* ***heard*** *military helicopters flying overhead.*

The form of the verb *to hear* in English spans across the area coloured in salmon pink in Figure 2.3, which covers these three functions. On the other hand, Maricopa *-'a* can express auditory as well as non-visual sensory evidentials, the light blue area. In this case, the region dedicated uniquely to auditory evidentials is justified by the opposition between the two languages.

2.1.2 The Inductive Bias in Children

Once established that cross-lingual universal tendencies stem from principles operating at the level of adult individuals, one may wonder if these are "*capacities that young humans share and presumably draw on in working out the structure of the language they hear*" (Bowerman, 2011, p. 1). In fact, Chomsky (1980) famously argued against the presumption that imitating behavioural patterns of adults alone is sufficient to learn a language, owing to the paucity of the stimuli available to children. Instead, it is necessary to postulate an inductive bias that accelerates learning either by pruning the search space of possible grammars or by favouring specific meaning-to-form mappings.

The first hypothesis has been formulated within generativism. In particular, this framework maintains that an innate component hard-wired in the brain constrains the possible formal grammatical structures. When exposed to a specific language, a child calibrates a set of binary 'parameters' (Lightfoot, 1979) that, for instance, determine the side in which dependents recursively fall with respect to their head, and whether specifiers and complements are on the same side (Graffi, 1980). While this conjecture explains the implicational universals in word order observed by Greenberg (1966), it fails to account for its exceptions. Moreover, the delay between the onset of linguistic production and the full command of a grammar in children was shown to remain stable irrespective of the flexibility of word order in the target language (Bowerman, 1973). Hence, the acquisition of formal structures appears to be mostly empirical.

On the other hand, there is ample evidence of a correspondence between universal tendencies in function-to-form mapping and language acquisition in children that accounts for their preparedness for language, as envisaged by the 'cognition hypothesis' (Slobin, 1973). First, this is corroborated by the so-called *emergent categories* (Clark, 2001), typical errors consisting in the over-extensions or under-extensions of lexical meaning that are not conventional in the target adult language but are quite common cross-lingually. These usually surface during the early stages of learning, but later vanish. For instance, the application of known words to new objects (e.g. *ball* for a PINCUSHION) is predominantly driven by shape similarity, the same criterion that informs numeral classifier systems across the world. In fact, both are arguably driven by principles common to human perception (Clark, 1976).

Moreover, several scientists (Bickerton, 2015; Slobin, 1985; Talmy, 1983) put forth the idea that grammatical meanings constitute the innate scaffold of semantics upon which learned content-word meanings are mounted. For instance, Slobin (1985) provides the example of the most salient temporal contrast, that is between results and processes. Children learning Turkish use the evidential past *-di/dɪ/ti/tɪ* with telic verbs (which

entail a completion), whereas they use the present tense *-iyor/ıyor/üyor/uyor* with atelic verbs (which entail duration). In general, this opposition transcends the formal means available in single languages like Turkish.

Although the degree to which core grammatical meanings are innate, as opposed to imitated from the target adult language, has been curtailed (Bowerman, 2011), it is safe to conclude that to some extent children are guided, in developing the knowledge of a language, by the same cognitive principles outlined in Section 2.1.1 and responsible for the typological universals in cross-lingual variation.

2.2 Machine Learning and Language

The acquisition of language in humans mentioned in Section 2.1.2 stands in stark contrast with the state-of-the-art practices in machine learning (Linzen, 2020). Ideally, probabilistic models should be able to imitate children's ability to fully command any new language from limited stimuli, in a sample-efficient fashion. In reality, the range of purposes of natural language is almost boundless: inner thought, social interactions, expression of emotions, search of information, creative performances, are just a few (Halliday, 1975). This list is both too wide to capture and too hard to evaluate quantitatively. Hence, machine learning usually addresses specific 'tasks'—rather than the language faculty as a whole—whose successful solution requires a certain degree of language knowledge and where the system performance is measurable. Moreover, the information available to machines is often purely textual, thus excluding grounding on other perceptual modalities (such as vision and speech) as well as communicative aspects of natural linguistic interactions (Bisk et al., 2020a).

2.2.1 Probability Theory

Textual linguistic data consist of variable-length sequences of discrete tokens (basic linguistic units such as words, characters, phonemes, etc). For a vocabulary of tokens $\{v \mid v \in V\}$, the Kleene closure defines the possible sequences $\{\mathbf{x} \mid \mathbf{x} \in V^\star\}$. The most fundamental task is *language modelling*, namely the assignment of a probability to any sequence in this set. This is an instance of *self-supervised* learning, which discovers the underlying structure from the data themselves without any additional guidance.

On the other hand, token sequences are often associated to labels from an inventory Y such that the constitute a dataset $\mathcal{D} \triangleq \{(\mathbf{x}, y) \mid \mathbf{x} \in V^\star, y \in Y\}$. This is an instance of *supervised* learning, as the goal is modelling the conditional probability of labels given

the corresponding sequences. The task is defined classification if labels are discrete, regression if they are continuous. Labels can also be themselves sequences. For instance, the task of POS tagging requires to tag each word with its part of speech; therefore, $\mathbf{y}_i \in \{\text{NOUN}, \text{VERB}, \dots\}^{|\mathbf{x}_i|}$.

For the sake of generality, both unlabelled and labelled data are defined here as observable events from a *sample space* $\mathcal{X}$. A *model* M is a set of probability measures on $\mathcal{X}$. Each measure is a function $p(\cdot)$ that maps from (sets of) events in the sample space to a real probability value such that $p : E \in \mathcal{X} \mapsto \mathbb{R}$ and it satisfies the 3 axioms of probability (Wasserman, 2013, p. 5): $p(E \in \Omega) \geq 0$, $p(\Omega) = 1$, and $p\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} p(E_i)$ if $\{E_1, E_2, \dots\}$ are disjoint. Each measure in the model M can be identified by its parameters $\boldsymbol{\vartheta}$ sampled from a parameter space $\mathcal{T} \in \mathbb{R}^d$, formally $M = \{m_{\boldsymbol{\vartheta}} \mid \boldsymbol{\vartheta} \in \mathcal{T}\}$. If the dimensionality of the parameters d is finite, the model is called parametric, and nonparametric otherwise.

The data $\mathcal{D} \triangleq \mathbf{x}_1(y_1), \dots, \mathbf{x}_n(y_n)$ from $\mathcal{X}$ are always observed in finite number and can be treated as random variables $\{X_1, X_2, \dots\}$. These are assumed to be sampled from a measure in the model m independently and identically distributed (i.i.d.):

$$X_1, \dots, X_n \sim_{i.i.d.} m_{\boldsymbol{\vartheta}}$$

The parameters are also a random variable Θ. From a Bayesian perspective, indeed, all sources of uncertainty are treated as random, including variables that are fixed but unknown. By making assumptions about the prior distribution $p(\Theta)$, the model becomes hierarchical. In particular, the observations are assumed to be generated from a two-step process (Orbanz, 2012):

$$\begin{aligned} \Theta &\sim p(\Theta) \\ X_1, \dots, X_n &\sim_{i.i.d} p(\cdot \mid \Theta) \triangleq M \end{aligned} \tag{2.6}$$

The relationship between random variables in Equation (2.6) can be expediently condensed in a graph, such as in Figure 2.4. Nodes represent random variables and are shaded if observed, clear if latent. Arrows correspond to the assumptions of dependency between variables. Finally, plates denote repetition of a variable of a specific kind.[7]

The parameters are often modelled as a multivariate Gaussian distribution, which is the maximum entropy distribution among those with support over $\mathbb{R}^d$, the space of $\mathcal{T}$,

[7]For brevity, I employ the value symbol in graphs and equations in lieu of the variable symbol: e.g., $\boldsymbol{\vartheta}$ in lieu of Θ.

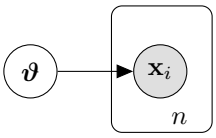


Figure 2.4 Graph of a minimal Bayesian generative model of the data.

and with only first and second moments defined: the mean $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{\vartheta})$ and the variance $\boldsymbol{\Sigma}_{ij} = \mathbb{E}((\boldsymbol{\vartheta}_i - \boldsymbol{\mu}_i)(\boldsymbol{\vartheta}_j - \boldsymbol{\mu}_j))$. This choice allows us to make as few assumptions as possible about the nature of the distribution. The probability density function of a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\vartheta} - \boldsymbol{\mu})\right) \tag{2.7}$$

The goal of machine learning is performing backward inference to estimate the distribution of the parameters given the observed data points. This amounts to calculating the posterior probability of Θ, which equals

$$\underbrace{p(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)}_{posterior} = \frac{\overbrace{\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\vartheta})}^{likelihood} \overbrace{\times\, p(\boldsymbol{\vartheta})}^{prior}}{\underbrace{\int \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\vartheta}) \times p(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}}_{evidence}} \tag{2.8}$$

according to Bayes' theorem (Bayes, 1763; Laplace, 1820). Modelling the likelihood $p(\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \boldsymbol{\vartheta})$ as $\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\vartheta})$ in Equation (2.8) relies on a crucial assumption: the parameters entirely contain the pattern underlying the observed data, and the remaining randomness decouples across samples (Orbanz, 2012). In particular, this is ensured in the second step of Equation (2.6): given Θ, each example is conditionally independent from the others, i.e. $X \coprod X' \mid \Theta$.

This assumption is viable if and only if the examples are *exchangeable*, according to De Finetti's theorem (De Finetti, 1929). Exchangeability implies that the order of observations is irrelevant. More formally, for any finite sequence of permutation of arbitrary pairs π of such order, the joint distribution of the examples remains unaffected, such that $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}) = p(\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(n+1)})$. Most remarkably, the assumption of symmetry inherent to exchangeability is essential for prediction (Zabell, 2005). In fact, under this condition the probability of a new example $\mathbf{x}_{n+1}$ can be inferred as:

$$p(\mathbf{x}_{n+1} \mid \boldsymbol{\vartheta}, \mathcal{D}) = \int p(\mathbf{x}_{n+1} \mid \boldsymbol{\vartheta}) \times p(\boldsymbol{\vartheta} \mid \mathcal{D})\, d\boldsymbol{\vartheta} \tag{2.9}$$

2.2.2 Artificial Neural Networks

The state-of-the art for natural language processing tasks is currently achieved by a precise brand of models, artificial neural networks (LeCun et al., 2015; Wang et al., 2019). In supervised settings, these models aim at learning a function from examples to labels $f_{\boldsymbol{\vartheta}} = X \mapsto \hat{Y}$ parameterised by $\boldsymbol{\vartheta} \in \mathbb{R}^d$.[8] This function is the composition of a series of sub-functions, the most fundamental one being an individual neuron. A neuron consists in an affine transformation of input array $\mathbf{x} \in \mathbb{R}^e$ into a single output $x' \in \mathbb{R}$ through a weight $\mathbf{w} \in \mathbb{R}^e$ and a bias $b \in \mathbb{R}$, followed by a non-linear deterministic function ϕ (called *activations*):

$$x' = \phi(\mathbf{w}^\top \mathbf{x} + b) \tag{2.10}$$

A list of activations ϕ relevant for this thesis, together with their corresponding derivatives with respect to the input, is presented in Table A.1. Multiple neurons can be juxtaposed together to constitute a *layer* with multiple outputs (whose number is the layer size h). In turn, layers can be stacked by feeding the output of the previous layer as the input to the next one. In particular, a 2-layer architecture where ϕ_1 is non-polynomial and $\phi_2 = \text{softmax}$, $h_1 \in \mathbb{N}$ and $h_2 = |Y|$ is a Multi-Layer Perceptron (MLP) classifier:

$$\hat{\mathbf{y}} = \text{MLP}(\mathbf{x}) = \text{softmax}(\mathbf{W}_2\, \phi(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) \tag{2.11}$$

The output $\hat{\mathbf{y}}$ represents $p(y \mid \mathbf{x}, f_{\boldsymbol{\vartheta},\boldsymbol{\alpha}})$, the conditional probability distribution of labels given a sentence under the feed-forward function. Note that f is uniquely characterised by the weight parameters, in this case $\boldsymbol{\vartheta} = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$, and the architecture parameters, in this case $\boldsymbol{\alpha} = \{\phi_1, h_1, \phi_2, h_2\}$. While the latter are usually treated as fixed hyper-parameters, the weights are the variable to be learned. Crucially, the function in Equation (2.11) is a universal approximator, meaning that it can represent any continuous function $f : \mathbb{R}^e \mapsto \mathbb{R}^{|Y|}$ for a suitable choice of parameters (Cybenko, 1989; Hornik, 1991).

How can $\boldsymbol{\vartheta}$ be learned then? First, if a correct label $y \in Y$ is observed, a *loss* function $\mathcal{L}(\hat{y}, y)$ can estimate the penalty for predicting $\hat{y}$ instead. In other words, this function measures the divergence between the predictive distribution $p(\hat{y} \mid \mathbf{x}, \boldsymbol{\vartheta})$ and

[8] Since d is finite, artificial neural networks are parametric models.

true label distribution $q \triangleq \delta_y$, where δ is the Kronecker delta. In particular, the average cross-entropy $\mathbb{H}(p, q)$ between discrete distributions over n examples is:

$$\mathcal{L}(\hat{Y}, Y) \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbb{H}(p^{(i)}, q^{(i)}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(-\sum_{y\in Y} p_y^{(i)} \ln q_y^{(i)}\right) \tag{2.12}$$

Thus learning is tantamount to finding the parameters that minimise the loss function in Equation (2.12): $\arg\min_{\boldsymbol{\vartheta}} \mathcal{L}(f_{\boldsymbol{\vartheta},\boldsymbol{\alpha}}(X), Y)$. Since the feed-forward function is non-convex due to the non-linear activations, however, no closed-form solution exists. Gradient descent provides an approximation that converges to a local minimiser by iteratively adjusting the parameters. For a learning rate η at the time step t:

$$\boldsymbol{\vartheta}_{t+1} = \boldsymbol{\vartheta}_t - \eta\nabla_{\boldsymbol{\vartheta}}\mathcal{L}(f_{\boldsymbol{\vartheta}_t,\boldsymbol{\alpha}}(X), Y) \tag{2.13}$$

The time complexity of a single step of Equation (2.13) is $\mathcal{O}(|\boldsymbol{\vartheta}|n)$ and depends on the number of data points n. In practice, this time is reduced by estimating the gradient stochastically on sub-samples of the data drawn uniformly at random, called batches $\mathcal{B} \subset \mathcal{D}$. This entails that the proxy gradient $\nabla_{\boldsymbol{\vartheta}}\mathcal{L}(f_{\boldsymbol{\vartheta}_t,\boldsymbol{\alpha}}(\mathcal{B}_X), \mathcal{B}_Y)$ is an unbiased estimator of the true gradient, retaining the same convergence guarantees (Peyré, 2020).[9] The gradient $\nabla_{\boldsymbol{\vartheta}}\mathcal{L}(\cdot)$ can be calculated efficiently through the back-propagation algorithm (Rumelhart et al., 1986).

In order to avoid over-fitting, it is customary to enforce an Occam's razor privileging specific parameter configurations. This is achieved by adding a *regulariser* $\mathcal{R}(\boldsymbol{\vartheta})$ to the cross-entropy function, such as an ℓ_2-norm of the parameter value. The relative importance of this second term is regulated by a hyper-parameter λ. As a result, the function to be minimised becomes $\mathcal{L}(\cdot) + \lambda\frac{1}{2}\sum_i \boldsymbol{\vartheta}_i^2$. Note that as $\lambda \to 0$, parameters are forced to decrease in magnitude, which encourages more complex, high-degree polynomial models. Vice versa, increasing λ favours simpler models.

Gazing through probabilistic spectacles, a neural network can receive the following interpretation (MacKay, 2003, pp. 492–495). Exponentiating the negative of the cross-entropy function in Equation (2.12), we obtain the definition of the likelihood $p(\mathcal{D} \mid \boldsymbol{\vartheta})$:

$$p(y \mid \mathbf{x}, \boldsymbol{\vartheta}, \boldsymbol{\alpha}) = \exp\left(-\mathcal{L}(f_{\boldsymbol{\vartheta},\boldsymbol{\alpha}}(\mathbf{x}), y)\right) \tag{2.14}$$

[9] To cancel out the noise introduced by sampling, $\eta \to 0$. Moreover, taking averages of previous iterations (*momentum*) into account can accelerate convergence.

Hence, the loss function is equivalent to the negative log-likelihood in Bayesian terms. Moreover, the ℓ_2 regulariser can be interpreted as a log-prior distribution over parameters, as follows:

$$p(\boldsymbol{\vartheta} \mid \lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\lambda \mathcal{R}(\boldsymbol{\vartheta})\right) \tag{2.15}$$

Hence, an ℓ_2 regulariser places a multivariate Gaussian prior $\mathcal{N}(\mathbf{0}, \frac{1}{\lambda} I)$ on the parameters. Combining the negative log-likelihood of Equation (2.14) and the log-prior of Equation (2.15), we recover the (unnormalised) posterior of Equation (2.8), $p(\boldsymbol{\vartheta} \mid \mathcal{D}, \lambda, \boldsymbol{\alpha}) \propto p(y \mid \mathbf{x}, \boldsymbol{\vartheta}, \boldsymbol{\alpha}) \times p(\boldsymbol{\vartheta} \mid \lambda)$. Rather than treating the parameters Θ as a random variable and inferring the full posterior, the gradient descent optimisation of Equation (2.13) only retrieves the point with *maximum a posteriori* (MAP) probability $\boldsymbol{\vartheta}^\star$, the (local) mode of the posterior distribution.[10] Thus, neural networks fail to account for the uncertainty surrounding the estimate of the parameters, staking everything on a single value.

This engenders at least two nefarious consequences. First, predictive inference in neural networks does not perform marginalisation over parameters of Equation (2.9), the hallmark of Bayesian learning. Instead, the (locally) optimal parameters are plugged into the equation such that $p(\mathbf{x}_{n+1} \mid \boldsymbol{\vartheta}, \mathcal{D}, \boldsymbol{\alpha}) = p(\mathbf{x}_{n+1} \mid \boldsymbol{\vartheta}^\star, \boldsymbol{\alpha})$. Second, when new data $\mathcal{D}_2$ become available after the MAP inference of $\boldsymbol{\vartheta}^\star$, the model needs to be re-trained from scratch lest to catastrophically forget the information encapsulated in the old data $\mathcal{D}_1$. Instead, after inferring the full posterior distribution one can simply continue learning through Bayesian updating (Nguyen et al., 2018):

$$p(\boldsymbol{\vartheta} \mid \mathcal{D}_1, \mathcal{D}_2, \boldsymbol{\alpha}) = \frac{p(\mathcal{D}_2 \mid \boldsymbol{\vartheta}, \mathcal{D}_1, \boldsymbol{\alpha}) \times p(\boldsymbol{\vartheta} \mid \mathcal{D}_1, \boldsymbol{\alpha})}{p(\mathcal{D}_2)} \tag{2.16}$$

2.2.3 Data Paucity

The over-confidence of MAP estimates is exacerbated in regimes of data paucity, when the number of observed examples is small. In the limit of infinite data, the posterior becomes peaked precisely at the MAP estimate:[11] $\lim_{n\to\infty} p(\boldsymbol{\vartheta} \mid \mathcal{D}) = \delta_{\boldsymbol{\vartheta}^\star}(\boldsymbol{\vartheta})$.[12] This is why MAP inference is called 'consistent', as it is guaranteed to find the correct value (or a likelihood-equivalent one, for non-identifiable models like neural networks), provided it lies in the hypothesis space.

[10] In absence of a regulariser, this becomes a *maximum likelihood* (ML) estimate.
[11] In the limit of infinite data, this is also the ML estimate, as the likelihood overwhelms the prior.
[12] The Dirac measure $\delta_x(A)$ equals 1 if $x \in A$ else 0.

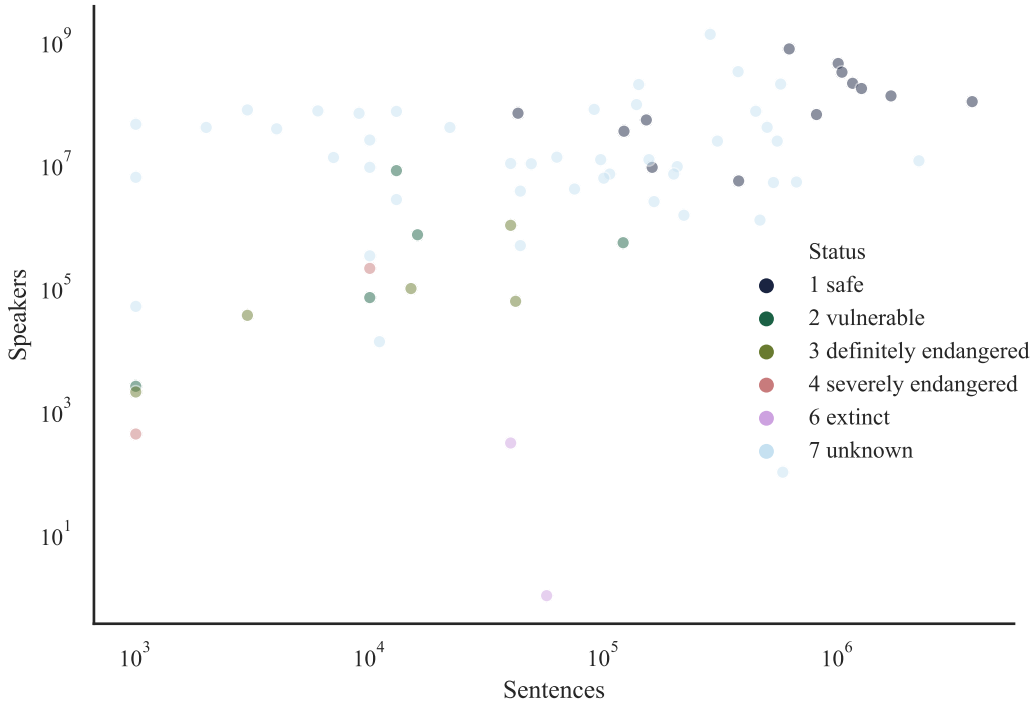


Figure 2.5 Number of speakers and sentences per language in Universal Dependencies 2.5, released in November 2019. The axes are in logarithmic scale.

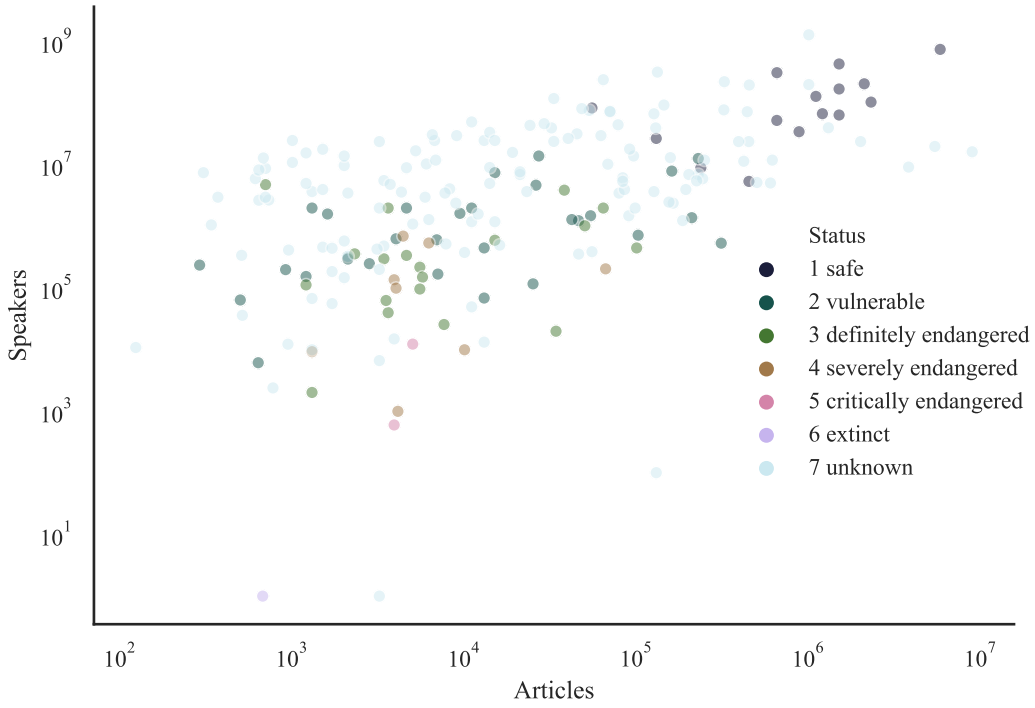


Figure 2.6 Number of speakers and articles per language in Wikipedia as of December 2018. The axes are in logarithmic scale.

In practice, however, the model has access only to finitely many observations. This raises a pivotal question: what is the influence of data paucity on the over-confidence of a neural model? To answer, we must turn the attention to the difference between the training error as formulated in Equation (2.12) and the generalisation error on new examples, as a function of the number of observed data points. While the training error can almost vanish in neural networks, to the extent that they can can memorise even random labels (Zhang et al., 2017), the generalisation error is also surprisingly moderate in practice. This holds true not despite—but exactly because—neural networks are broadly over-specified, as $|\mathcal{D}| \ll |\vartheta|$ (Bartlett, 1998). A classical theorem on the bound between the training loss given the (locally) optimal function $\mathcal{L}_{f^\star_\vartheta}$ and the generalisation loss given the true function $\mathcal{L}_{f_\vartheta}$ states that for any family of functions $\mathcal{F}$, with probability $1-\delta$ (Mohri et al., 2018, Theorem 3.1 p. 35):

$$\mathcal{L}_{f_{\hat{\vartheta}}} - \mathcal{L}_{f_\vartheta} \geq 2\,\mathfrak{R}_{\mathcal{D}}(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{|\mathcal{D}|}} \tag{2.17}$$

where $\mathfrak{R}_{\mathcal{D}}(\mathcal{F})$ is the empirical Rademacher complexity with respect to a family of functions $\mathcal{F}$ and a data sample $\mathcal{D}$:

$$\mathfrak{R}_{\mathcal{D}}(\mathcal{F}) = \mathop{\mathbb{E}}_{\xi\sim\{\pm1\}^{|\mathcal{D}|}}\left[\sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}\xi f(\mathbf{x})\right] \tag{2.18}$$

The Rademacher complexity in Equation (2.18) for a network like the one described in Equation (2.11) decreases with a scale proportional to 1) the width of the hidden layers—hence, the number of parameters— (Neyshabur et al., 2019)[13] and 2) the square root of the training data size $\sqrt{|\mathcal{D}|}$. The same growth rate is followed by the second term in the right-hand side of Equation (2.17). Crucially, this implies that for small amounts of data points, the gap between the two losses spreads. As a consequence, ML estimation becomes unreliable (Bottou and Bousquet, 2008).

Crucially, the provision of an abundance of labelled linguistic data does not only run counter to the natural learning process of human children (see Section 2.1.2), but also violates the constraints imposed by data availability for most combinations of languages and tasks. In fact, the creation of annotated data is time-consuming and skill-intensive. Hence, most of the linguistic resources are oligopolised by a handful of well-researched languages. What is more, also unlabelled data are concentrated in a few languages

[13]Note that the proof only accounts for 2-layer MLPs with a ReLU activation.

whose communities of speakers have a substantial presence on the internet, leading to the so-called 'digital divide' (Kornai, 2013).[14]

To make some concrete examples, let us consider the resources that currently cover the broadest set of languages. The Universal Dependencies treebanks contain sentences in 90 languages annotated for dependency parsing, whereas Wikipedia dumps contain unlabelled texts for 278 languages.[15] Not only this number pales in comparison to the total world's languages, but the data distribution per language is highly imbalanced, as shown in Figure 2.5 for Universal Dependencies sentences and in Figure 2.6 for Wikipedia articles. Crossing this information with the number of native speakers and with the UNESCO language status (from 1 safe to 6 extinct),[16] it emerges that resource-poor languages tend to be more endangered, but their speaker communities may be as large as those of resource-rich languages. This hints at the fact that available data is not only imbalanced in terms of number of data points, but also not representative of the language variation across the world.

2.2.4 Cross-lingual Knowledge Transfer

Data paucity poses an insurmountable obstacle to supervised learning in resource-poor languages, as well as to inherently multilingual applications such as machine translation (Artetxe et al., 2018, 2019; Lample et al., 2018a,b). Nevertheless, the knowledge needed to solve a task may be available through data in other domains. In fact, such knowledge can be transferred from other tasks, assuming that the required sets of skills are synergic and partly overlapping (Ruder et al., 2019a), or from other modalities such as vision (Lu et al., 2019, *inter alia*). The most widespread and effective source of transferable knowledge, however, are other resource-rich languages. In fact, as argued in Section 2.1, while languages vary formally owing to the arbitrariness of the sign, semantic functions are universal. Obviously, the main challenge in this case is that source language s and target language t do not share their sample spaces, hence $X_s \neq X_t$. This can be overcome by transferring either the annotation to the other space or the model parameters.

Annotation transfer relies on projecting the labels from a source text to a target parallel text, as pioneered by Yarowsky et al. (2001) and Hwa et al. (2005). If the

[14] In fact, 34% of the world's languages are not even recorded in writing despite their status being vigorous (Lewis et al., 2016). As of March 2015, just 40 out of the 188 languages documented on the internet accounted for 99.99% of the web pages, according to https://w3techs.com/technologies/overview/content_language/all.

[15] Universal Dependencies version 2.5 was released on 15 November 2019. Per language article counts of Wikipedia dumps were last updated on December 2018.

[16] This information was queried from Wikidata on 8 April 2020.

annotation is token-level, it further requires to word-align the parallel texts before the projection. Afterwards, a model can be trained through supervised learning on the resulting target annotation. This approach, however, is hampered by several factors. First, creating word-alignment systems demands parallel texts in the first place. Second, errors inherent to such systems pile up along the projection pipeline (Agić et al., 2015). Third, and most importantly, token-level projection assumes that all linguistic structures are preserved in translation. However, this is patently false, due to typological variation. While post-processing can partially amend this discrepancy, by filtering out annotations that are infrequent or with low confidence (Padó and Lapata, 2009), this would also introduce unwanted bias into the training data.

An alternative method for cross-lingual transfer is *translating* the target language into English during evaluation (Conneau et al., 2018; Lewis et al., 2019). Thus, an English model can be deployed on top of the resulting data. Otherwise, English can be translated into the target language before training. This is achieved through a machine translation (MT) model (Banea et al., 2008) or a bilingual lexicon (Durrett et al., 2012). The annotation is then projected and used to supervise training in the target language. Between translating evaluation or training data, the former is by far the more successful (Conneau et al., 2018; Lewis et al., 2019). However, both assume the availability of reliable MT systems, and again cross-lingual isomorphism in linguistic structure.

Model transfer, finally, offers higher flexibility (Conneau et al., 2018) and involves training a model directly on the source data and deploying it onto target data (Zeman and Resnik, 2008). This entails mapping both source and target data onto a language-agnostic representation, for instance part-of-speech tags and morphological features (Zhang et al., 2012), or multilingual Brown word clusters (Täckström et al., 2012). This approach, however, reaches it full potential by mapping linguistic units from multiple languages into distributed representations in a shared space $f : t \rightarrow \mathbf{x} \in \mathbb{R}^e$ inferred through unsupervised learning, known as *word embeddings*.

Static word embeddings, identical for any instance of a token throughout a text, are inspired by the distributional hypothesis (Firth, 1957; Harris, 1951) and are pre-trained based on word co-occurrence information in corpora (Bojanowski et al., 2017; Ruder et al., 2019b; Upadhyay et al., 2016). On the other hand, contextualised word embeddings assign a representation to each token dependent on its surroundings, providing a proxy for meaning in context, by pre-training an encoder network through language modelling (Conneau et al., 2020; Conneau and Lample, 2019; Devlin et al., 2019, *inter alia*). By said method, given some raw texts in both the source and target languages $\{\mathbf{x} \mid \mathbf{x} \in \mathcal{D}_s \cup \mathcal{D}_t\}$, the parameters of an encoder $\boldsymbol{\vartheta}_{\text{ENC}}$ are first optimised. Subsequently, a classifier with

parameters $\boldsymbol{\vartheta}_{\text{CLS}}$ (usually a feed-forward network such as an MLP) is stacked on top of the encoder, and randomly initialised. Both are jointly 'fine-tuned' through labelled data in the source language $\{(\mathbf{x}, y) \mid (\mathbf{x}, y) \in \mathcal{D}_s\}$. Incidentally, this method achieves the state of the art on multilingual benchmarks across assorted tasks (Hu et al., 2020), but receives exposure to a number of examples multiple orders of magnitude higher than children (see the poverty of the stimulus argument from Section 2.1.2). In fact, explicit guidance from adults is negligible compared to the profuse and entangled information supplied by raw perceptual inputs (Gorder, 2006, quoting Geoffrey Hinton in 1996).

Given a target resource-poor language, the *selection* of the most suited source language(s) among those with labelled data is no less paramount than the method of transfer. Originally, the choice was driven by similarity in formal structures, devising metrics based on typological features (Deri and Knight, 2016), part-of-speech tag distributions (Rosa and Zabokrtsky, 2015), or dependency tree edit distance (Ponti et al., 2018a). This is predicated on the assumption that narrowing the gap between sample spaces X_s and X_t facilitates transfer. Obviously, there may be a trade-off between language similarity and the abundance of data in candidate source languages. In fact, with the advent of unsupervised pretraining, it is the latter, and hence the reliability of the shared representation, that plays a pivotal role in the success of cross-lingual transfer (Lauscher et al., 2020).

After fine-tuning the model on the source language, the transferred model can either perform predictive inference on target examples directly, a setting known as *zero-shot learning*, or be further updated on a small number of target labelled data, known as *few-shot learning*. Although successful, multilingual pre-training and fine-tuning are inadequate in light of the discussion so far:

- Since initialising parameters obtained through maximum likelihood estimates may incur catastrophic forgetting (see Section 2.2.2), the final model risks to lose the memory of both unlabelled data and source annotated data in between the two transfer steps.

- The focus on formally similar languages in source selection infringes a key finding of human language acquisition (see Section 2.1): namely, that the grammatical knowledge of children is not biased towards specific formal strategies, but rather mirrors the universal patterns in meaning-to-form mapping.

- Massively multilingual pretraining is riddled by the 'curse of multilinguality' (Cao et al., 2020; Conneau et al., 2020; Hu et al., 2020): the more languages covered,

the more performance collapses, as the model needs to 'cram' information about multiple separate data distributions into a single set of parameters.

In this thesis, I argue that all these limitations can be solved elegantly by recasting neural knowledge transfer into the framework of Bayesian learning. In particular, rather than leveraging formal similarities, I seek to recreate a language-universal inductive bias. Since the cognitive and communicative principles that guide learning in children are inaccessible without grounding, however, I simulate their effect by 'reverse-engineering' a representative sample of source languages. Contrary to point estimates, Bayesian inference allows for taking uncertainty into account. Thus, a prior distribution over neural models of language (i.e. over both neural weights and architectures) would mirror the variation across possible languages. This prior can then be harnessed to accelerate the process of learning a target language in zero-shot and few-shot settings.

2.3 Bayesian Neural Models

In order to recast cross-lingual neural transfer into a Bayesian framework, it is necessary to ask the same questions explored in Section 2.2.2 for supervised neural learning. Under which conditions is a model expected to generalise to new languages? The answer relies again on de Finetti's theorem: generalisation rests on the assumption of symmetry across examples, which is obviously false for data from a source language and a target language. In fact, the respective joint distributions are different regardless of the sample size, i.e. $p(\mathcal{D}_s) \neq p(\mathcal{D}_t)$. Hence, the model defined by Figure 2.4 does not provide sufficient guarantees for generalisation. Instead, one has to posit that language-specific parameters are, in turn, exchangeable and hence conditionally independent given a higher-order variable Φ:

$$\begin{aligned} \Phi &\sim p(\Phi) \\ \Theta_1, \ldots, \Theta_m &\sim_{i.i.d} p(\cdot \mid \Phi) \\ X_1, \ldots, X_n &\sim_{i.i.d} p(\cdot \mid \Theta_i) \end{aligned} \tag{2.19}$$

This translates into the *hierarchical* graphical model of Figure 2.7. Therefore, cross-lingual transfer can be thought as seeking a language-universal prior $\boldsymbol{\varphi}$. Continuing the analogy with language acquisition in children established in Section 2.1, this prior should capture all aspects of cognition that precede and accompany the experience of linguistic stimuli.

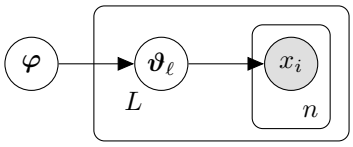


Figure 2.7 Graph of a hierarchical Bayesian generative model.

This perspective can shed light on the widespread approach where a point estimate $\boldsymbol{\vartheta}^\star$ from 'pre-training' serves as initialisation for 'fine-tuning' on target data $\mathcal{D}_t$ through gradient descent as in Equation (2.13), and reveal that it posits an implicit prior. In fact, Santos (1996) proved that for linear functions, truncated optimisation starting from a specific initialisation is equivalent to performing maximum-a-posteriori inference on the model in Figure 2.7 where $\boldsymbol{\varphi}$ is a fixed prior $\mathcal{N}(\boldsymbol{\vartheta}^\star, \boldsymbol{Q})$ and $\boldsymbol{Q}$ depends on the learning rate η, the step number t, and the co-variance matrix of $\mathcal{D}_s$. The implicit objective that is maximised by performing gradient descent from initialisation $\boldsymbol{\vartheta}^\star$ is then:

$$\arg\max_{\boldsymbol{\vartheta}} p(\mathcal{D}_t \mid \boldsymbol{\vartheta}) - (\boldsymbol{\vartheta}^\star - \boldsymbol{\vartheta})^\top \boldsymbol{Q}^{-1} (\boldsymbol{\vartheta}^\star - \boldsymbol{\vartheta}) \tag{2.20}$$

Instead of leaving the variance $\boldsymbol{Q}$ implicit in the optimisation procedure, we could improve the prior by estimating it explicitly from pre-training. How to perform Bayesian inference of the distribution over neural parameters given source data then, rather than a mere point estimate $\boldsymbol{\vartheta}^\star$? As it is obvious from Equation (2.8), treating Θ as a latent variable would require to integrate over all possible $\boldsymbol{\vartheta}$ values, which is utterly intractable. Since an exact solution is off-limits, an approximation is in order.

A first set of Bayesian approximations proposed for neural networks is characterised by discrete support, as they assign zero probability mass almost everywhere in the neural parameter space (Wilson, 2019). These include Monte-Carlo Dropout, where the expectation over a posterior distribution is approximated as the average of a series of feed-forward passes with different dropout patterns (Gal and Ghahramani, 2016a); Deep Ensembles, where multiple neural networks are combined and their predictive distribution is smoothed through adversarial training (Lakshminarayanan et al., 2017); and Stochastic Gradient Lengevin Dynamics (Welling and Teh, 2011). However, because of their discrete support, these methods provide highly skewed distributions: when performing Bayesian updates, if the true solution is found outside the few probability 'spikes', they cannot converge. In other words, no amount of data can overwhelm such prior.

2.3.1 Deterministic Inference

Methods for Bayesian inference with continuous support can be divided into two families. Recall that the need of approximation stems from the presence of latent variables, whose posterior cannot be computed analytically. *Deterministic* methods, such as Laplace (MacKay, 1992) and variational approximations (Blundell et al., 2015) bound the effects of introducing such variables, and optimise this bound. Instead, *imputation* methods require to sample from the latent variable—for instance, though Monte Carlo sampling (Neal, 1996)—and condition dependent variables on such value. Because of their ability to scale seamlessly to large models, throughout the current thesis I will take into consideration only deterministic methods, which I outline below.

Laplace Approximation

The Laplace method simply approximates the true (possibly multi-modal and non-Gaussian) probability of the neural weights with a multi-variate Gaussian, whose parameters have to be determined. Assume the mode of the (unnormalised) probability density $p^\star(\boldsymbol{\vartheta} \mid \mathcal{D})$ is known; for instance, it can be obtained in neural networks as the MAP estimate $\arg\max_{\boldsymbol{\vartheta}} \mathcal{L}(f_{\boldsymbol{\vartheta}}(X), Y) + \lambda\mathcal{R}(\boldsymbol{\vartheta})$. Afterwards, it is sufficient to Taylor-expand the unnormalised log-probability around such peak value (MacKay, 2003, p. 341):

$$\log p^\star(\boldsymbol{\vartheta} \mid \mathcal{D}) \approx \log p^\star(\boldsymbol{\vartheta}^\star \mid \mathcal{D}) - \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \mathbf{H} (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star) + \dots \tag{2.21}$$

where $\mathbf{H}$ is the Hessian, the matrix of second order derivatives of the log-probability with respect to the parameters evaluated at the mode:

$$\mathbf{H}_{ij} = \frac{\delta^2}{\delta_{\boldsymbol{\vartheta}_i} \delta_{\boldsymbol{\vartheta}_j}} \log p^\star(\boldsymbol{\vartheta} \mid \mathcal{D}) \Big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^\star} \tag{2.22}$$

Note that the first-order term of the Taylor expansion $(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \nabla p^\star(\mathcal{D} \mid \boldsymbol{\vartheta}^\star)$ in Equation (2.21) is dropped as by definition the gradient at the optimum is zero. Then the posterior $p(\boldsymbol{\vartheta} \mid \mathcal{D})$ can be approximated by plugging the approximation based on Taylor expansion of Equation (2.21) into Bayes theorem of Equation (3.3):

$$p(\boldsymbol{\vartheta} \mid \mathcal{D}) \approx \frac{\exp\left[p^\star(\boldsymbol{\vartheta}^\star \mid \mathcal{D}) + \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \mathbf{H} \, (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)\right]}{\int \exp\left[p^\star(\boldsymbol{\vartheta}^\star \mid \mathcal{D}) + \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \mathbf{H} \, (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)\right] \mathrm{d}\boldsymbol{\vartheta}} \tag{2.23}$$

By simplifying the term $p^\star(\boldsymbol{\vartheta}^\star \mid \mathcal{D})$ and evaluating the integral, we obtain:

$$\frac{\exp\left[-\frac{1}{2}(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^{\star})^{\top}(-\mathbf{H})(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^{\star})\right]}{\sqrt{(2\pi)^{d}\left|-\mathbf{H}\right|^{-1}}} \triangleq \mathcal{N}(\boldsymbol{\vartheta}^{\star}, -\mathbf{H}^{-1}) \tag{2.24}$$

In other words, the Laplace method approximates a posterior distribution by a Gaussian whose mean is the MAP estimate $\boldsymbol{\vartheta}^{\star}$ and whose co-variance is the negative inverse of the Hessian, $-\mathbf{H}^{-1}$. This method is employed in the experiments of Chapter 3, where I further elaborate on its implementation.

Variational Inference

Variational approximation is an alternative inference technique to deal with the intractable integral arising in Equation (2.8). In particular, it consists in deriving a lower bound of the log-evidence of the data, called ELBO, by introducing a surrogate distribution over weights $q(\boldsymbol{\vartheta}) \triangleq \mathcal{N}(\boldsymbol{\vartheta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and minimising its Kullbach-Leibler divergence from the true posterior $p(\boldsymbol{\vartheta} \mid \mathcal{D})$:

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathbb{KL}\left[q(\boldsymbol{\vartheta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid\mid p(\boldsymbol{\vartheta} \mid \mathcal{D})\right] \tag{2.25}$$

Expanding Equation (2.25) by the definition of Kullbach-Leibler divergence, one obtains:

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \int q(\boldsymbol{\vartheta}) \log \frac{q(\boldsymbol{\vartheta})}{p(\mathcal{D} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta})} \, \mathrm{d}\boldsymbol{\vartheta} \tag{2.26}$$

Note that $p(\mathcal{D})$ disappears because it shares no arguments with the arg min operator. Then by the definition of expectation the objective becomes:

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathbb{KL}\left[q(\boldsymbol{\vartheta}) \mid\mid p(\boldsymbol{\vartheta})\right] - \mathbb{E}_{q(\boldsymbol{\vartheta})} \log p(\mathcal{D} \mid \boldsymbol{\vartheta}) \tag{2.27}$$

where the first term (known as complexity cost) is the divergence between the learned posterior and the prior over weights, and the second term is the familiar log-likelihood. In practice, the complexity cost has a closed-form solution if the prior is also multivariate Gaussian. Blundell et al. (2015) proposed an algorithm, called Bayes by Backprop, to optimise Equation (2.27) for neural network weights. Under mild assumptions, it can be shown that the gradient of the expectation in Equation (2.27) is equivalent to the expectation of the gradient. Furthermore, the weights $\boldsymbol{\vartheta}$ can be Monte Carlo sampled via a deterministic function (known as reparametrisation trick) given the Gaussian parameters and some noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as $\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\odot$ stands for element-wise multiplication. Thus, averaging over repeated samples returns an unbiased estimate of

the gradients. As a consequence, the neural network can be trained through regular gradient descent via back-propagation.

In addition to enabling end-to-end learning during posterior inference, the reparametrisation trick comes in handy also for predictive inference. In fact, it allows to approximate model averaging, the integration over weights appearing in Equation (2.9), as an average over samples. Although the performance is not always superior to simply plugging in the learned mean, (approximate) model averaging yields smoother distributions, which better quantify uncertainty in predictions. I resort to variational inference for such purpose in the experiments in Chapter 5, where I derive distinct ELBOs for the specific model proposed therein.

2.3.2 Empirical Bayes

Bayesian inference is not limited to neural weights. In fact, anything measurable can be modelled as a variable under this framework. This includes, for instance, the architecture of neural networks. Rather than fixed hyper-parameter selected through grid search, parameters such as layer width and depth, or the choice of non-linear activation functions, can be learned during the training phase. Assuming the conditional independence of the data from the architecture given the weights, the objective becomes:

$$\arg\max_{\boldsymbol{\alpha}} \int p(\mathbf{x} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\vartheta} \tag{2.28}$$

where an intractable integral over weights resurfaces. A cheap and naive solution is Empirical Bayes, which estimates the maximum likelihood value of the architecture first, and then freezes it during the weight optimisation. This approach, however, is prone to over-fitting and over-confidence in predictive inference (Murphy, 2012, p. 173). I propose more sophisticated inference schemes for hierarchical Bayesian models as defined by Equation (2.28) in Chapter 4.

2.4 Summary

The remarkable facts that cross-lingual variation is bounded rather than random, and that children can pick up languages with limited stimuli, both find a common justification. In fact, language is grounded on real-world communication and embodied in human perception and cognition. These constrain the paths of language innovation and guide language acquisition. In such a way, they provide an inductive bias for learning and enable generalisation to new usages of language.

While machine learning has made great strides in natural language processing, the current state-of-the-art approaches—pretraining an encoder through self-supervision on language modelling, and subsequently fine-tuning it on few annotated examples—struggle with scenarios characterised by data paucity or distribution shift. Since this hinders their effectiveness in most non-trivial applications and for most of the world's languages, it is crucial to adopt solutions to correct these limitations.

In this thesis, I propose that this can be achieved by re-aligning machine learning with some desirable properties of language acquisition in humans within a Bayesian framework. In particular, I supply neural functions with an inductive bias by constructing a prior (encompassing both weights and architectures) through cross-lingual knowledge transfer and hand-crafted typological features. This requires to perform inference through deterministic techniques such as the Laplace and variational approximations. Moreover, I explore the potential of graphical models to express the structured nature of linguistic knowledge, and leverage this factorisation of the space of neural weights to improve generalisation.

Chapter 

A Prior over Weights for Language Modelling

3.1 Introduction

Despite their success in core natural language processing tasks, neural networks remain black-box models: it is arduous to interpret if they capture linguistic knowledge or rather spurious correlations while solving a task. In turn, such propensity towards 'true' linguistic knowledge is desirable as it arguably leads to the same sort of generalisation as humans. In particular, researchers have turned their attention to *probing* the 'inductive bias' towards language (Linzen et al., 2016; Marvin and Linzen, 2018; Ravfogel et al., 2018), intended as the abstractions a model is capable of, exhibited by manually designed architectures such as LSTMs (Hochreiter and Schmidhuber, 1997). For example, do they learn syntax (Marvin and Linzen, 2018)? Do they map onto grammaticality judgements (Warstadt et al., 2019)? The focus on architectures stems from the fact that hyper-parameters $\boldsymbol{\alpha}$ constrain the family of functions that a neural network can learn (see Section 2.2.2). This recent vein of research, however, implicitly assumes a uniform (unnormalizable) prior over the space of neural weight parameters $\boldsymbol{\vartheta}$ (Ravfogel et al., 2019, *inter alia*).

In this chapter of my thesis, in contrast, I aim at *providing* the correct inductive bias by finding a prior distribution over network parameters that generalise well for human language. Thus, the focus is not only on the *content* of the inductive bias, but also on its ability to achieve *sample efficiency*. In particular, I take a Bayesian-updating approach and approximate the posterior distribution over the network parameters conditioned on the data from a sample of *seen* training languages using the Laplace method (Azevedo-

Filho and Shachter, 1994; MacKay, 1992).[1] Afterwards, such distribution serves as a prior for maximum-a-posteriori (MAP) estimation of network parameters for each held-out *unseen* language.

This study focuses on the task of character-level language modelling (Cotterell et al., 2018; Gerz et al., 2018b; Mielke et al., 2019). In particular, I opt for an open-vocabulary setup where all tokens—including infrequent ones—are preserved rather than being substituted with a special <UNK> token. This is crucial for a fair comparison of model performances across languages (Gerz et al., 2018a). Taking characters as a proxy for the underlying phonemes, the search for a universal prior is then motivated by the notion that knowledge about likely phoneme inventories and their combinations precedes experience in humans, as argued in Section 2.1.2. Although the nature of such prior knowledge is still disputed—innate constraints whose ranking is language-specific for generativist frameworks like Optimality Theory (Smolensky and Prince, 1993), or the outcome of general articulatory and perceptual principles for functionalist frameworks—at least two key aspects are unanimously agreed upon, which are relevant to this experiment: i) this knowledge facilitates language acquisition (Chomsky, 1959); and ii) it is reflected in cross-lingual variation (Gilligan, 1989; Graffi, 1980), which implies that it can be reconstructed from it to some extent.

In this chapter, I investigate whether a suitable prior over weight parameters can encapsulate such prior knowledge, and make machines achieve sample efficiency in learning. In particular, I run experiments under several regimes of data scarcity for the held-out languages (zero-shot, few-shot, and joint multilingual learning) over a sample of 77 typologically diverse languages.

As an orthogonal contribution, I also explore a regime where the universal prior is conditioned on *side information* from typology. Realistically, in fact, a model should not be completely in the dark about held-out languages, as coarse-grained features about general linguistic properties are documented for most of the world's languages and available in typological databases such as URIEL (Littell et al., 2017), as mentioned in Section 2.1.1. In particular, I consider several techniques for conditional language modelling from the literature, including: i) concatenating typological features to hidden states (Östling and Tiedemann, 2017) and ii) generating the weight parameters through hyper-networks receiving typological features in input (Platanios et al., 2018).

Empirically, given the results of this study, I offer two findings. The first is that neural recurrent models with a universal prior significantly outperform baselines with

[1] In principle, alternative inference schemes such as variational inference (Blundell et al., 2015) or Hamiltonian Monte Carlo (Neal, 2011) could serve the same purpose.

uninformative and unnormalisable priors both in zero-shot and few-shot training settings. Secondly, conditioning on typological features further reduces the test error in the few-shot setting, but I report negative results for the zero-shot setting, possibly due to some inherent limitations of typological databases (Ponti et al., 2019a).

The study of low-resource language modelling also holds promise to have a benign impact on society. As shown in Section 2.2.3, the digital footprint of most of the world's languages is almost insignificant. What is more, Kornai (2013) prognosticates that the digital divide will act as a catalyst for the extinction of many of the world's languages. The transfer of language technology may help reverse this course and give access to vital services to unrepresented communities of speakers. My work is a step in this direction, given that character-level language modelling lies at the core of tasks such as text-to-speak and morphological analysis. What is more, because of its generality, the proposed method is amenable to be deployed in other natural language processing applications in the future.

3.2 LSTM Language Models

In this work, I address the task of *character-level* language modelling. Whereas word lexicalization is mostly arbitrary across languages, phonemes allow for transferring universal constraints on phonotactics[2] and language-specific sequences that may be shared across languages, such as borrowings and cognates (Brown et al., 2008). Since languages are mostly recorded in text rather than phonemic symbols (IPA), however, I focus on characters as a loose approximation of phonemes.

Let Σ_ℓ be the set of characters for language ℓ. Moreover, consider a collection of languages $\mathcal{T} \sqcup \mathcal{E}$ partitioned into two disjoint sets of observed (training) languages $\mathcal{T}$ and held-out (evaluation) languages $\mathcal{E}$. Then, let $\Sigma = \cup_{\ell \in (\mathcal{T} \sqcup \mathcal{E})} \Sigma_\ell$ be the union of character sets in all languages. A universal, character-level language model is a probability distribution over Σ^*.[3] Let $\mathbf{x} \in \Sigma^*$ be a sequence of characters. We write:

$$p(\mathbf{x} \mid \boldsymbol{\vartheta}) = \prod_t p(x_t \mid \mathbf{x}_{<t}, \boldsymbol{\vartheta}) \tag{3.1}$$

where t is a time step, $\boldsymbol{\vartheta}$ are the parameters, and every sequence $\mathbf{x}$ starts (ends) with a distinguished start-of-sentence (end-of-sentence) symbol.

[2]E.g. with few exceptions (Evans and Levinson, 2009, sec. 2.2.2), the basic syllabic structure is vowel–consonant.

[3]Note that Σ is also augmented with punctuation and white space, and distinguished beginning-of-sequence and end-of-sequence symbols, respectively.

We implement character-level language models with Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). These encode the entire history $\mathbf{x}_{<t}$ as a fixed-length vector $\mathbf{h}_t$ by manipulating a memory cell $\mathbf{c}_t$ through a set of gates. Then I define

$$p(x_t \mid \mathbf{x}_{<t}, \boldsymbol{\vartheta}) = \text{softmax}(\mathbf{W}\,\mathbf{h}_t + \mathbf{b}). \tag{3.2}$$

LSTMs have an advantage over other recurrent architectures as memory gating mitigates the problem of vanishing gradients and captures long-distance dependencies (Pascanu et al., 2013).

3.3 Neural Language Modelling with a Universal Prior

The fundamental hypothesis of this work is that there exists a prior $p(\boldsymbol{\vartheta})$ over the weights of a neural language model that places high probability on networks that describe human-like languages. Such a prior would provide an inductive bias that facilitates learning *unseen* languages. In practice, I construct it as the posterior distribution over the weights of a language model of *seen* languages. Let $\mathcal{D}_\ell$ be the examples in language ℓ, and let $\mathcal{D}_\mathcal{T}$ be the examples in all training languages $\cup_{\ell=1}^{|\mathcal{T}|}\mathcal{D}_\ell$. Taking a Bayesian approach, the posterior over weights is given by Bayes' rule:

$$\underbrace{p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T})}_{\textit{posterior}} \propto \underbrace{\prod_{\ell\in\mathcal{T}} p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta})}_{\textit{likelihood}}\ \underbrace{p(\boldsymbol{\vartheta})}_{\textit{prior}} \tag{3.3}$$

I take the prior of Equation (3.3) to be a Gaussian with zero mean and covariance matrix $\sigma^2\,\mathbf{I}$, i.e.

$$p(\boldsymbol{\vartheta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}||\boldsymbol{\vartheta}||_2^2\right). \tag{3.4}$$

However, computation of the posterior $p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T})$ is woefully intractable: recall that, in our setting, each $p(\mathcal{D}_\mathcal{T} \mid \boldsymbol{\vartheta})$ is an LSTM language model, like the one defined in Equation (3.2). Hence, I opt for a simple approximation of the posterior, using the classic Laplace method (MacKay, 1992). This method has recently been applied to other transfer learning or continuous learning scenarios in the neural network literature (Kirkpatrick et al., 2017; Kochurov et al., 2018; Ritter et al., 2018).

In Section 3.3.1, I first introduce the Laplace method, which approximates the posterior with a Gaussian centred at the maximum-likelihood estimate.[4] Its covariance matrix is amenable to be computed with backpropagation, as detailed in Section 3.3.2. Finally, I describe how to use this distribution as a prior to perform maximum-a-posteriori inference over new data in Section 3.3.3.

3.3.1 Laplace Method

First, I (locally) maximise the logarithm of the RHS of Equation (3.3):

$$\mathcal{L}(\boldsymbol{\vartheta}) = \sum_{\ell \in \mathcal{T}} \log p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}) + \log p(\boldsymbol{\vartheta}) \tag{3.5}$$

We note that this is equivalent to the log-posterior up to an additive constant, i.e.

$$\log p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T}) = \mathcal{L}(\boldsymbol{\vartheta}) - \log p(\mathcal{D}_\mathcal{T}) \tag{3.6}$$

where the constant $\log p(\mathcal{D}_\mathcal{T})$ is the log-normalizer. Let $\boldsymbol{\vartheta}^\star$ be a local maximizer of $\mathcal{L}(\boldsymbol{\vartheta})$.[5] We now approximate the log-posterior with a second-order Taylor expansion around $\boldsymbol{\vartheta}^\star$:

$$\log p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T}) = \mathcal{L}(\boldsymbol{\vartheta}^\star) + \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \mathbf{H} \, (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star) + R - \log p(\mathcal{D}_\mathcal{T}) \tag{3.7}$$

where $\mathbf{H}$ is the Hessian matrix and R are higher-order terms. Note that I have omitted the first-order term, since the gradient $\nabla_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}) = 0$ at the local maximizer $\boldsymbol{\vartheta}^\star$. This quadratic approximation to the log-posterior is Gaussian, which can be seen by exponentiating both sides in Equation (3.7):

$$p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T}) \propto \exp\left(\frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \mathbf{H} \, (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)\right) \tag{3.8}$$

where $\exp(\mathcal{L}(\boldsymbol{\vartheta}^\star))$ is absorbed into the Gaussian's normalisation constant, which may be computed analytically. Because $\boldsymbol{\vartheta}^\star$ is a local maximizer, $\mathbf{H}$ is a negative semi-definite matrix.[6] In principle, computing the Hessian is possible by running backpropagation

[4] Note that, in general, the true posterior is multi-modal. The Laplace method instead approximates it with a unimodal distribution.

[5] In practice, non-convex optimization is only guaranteed to reach a critical point, which could be a saddle point. However, the derivation of Laplace's method assumes that we do reach a maximizer.

[6] Note that, as a result, our representation of the Gaussian is non-standard; generally, the precision matrix is positive semi-definite and is accompanied by a negative sign.

twice: this yields a matrix with d^2 entries. However, in practice, this is impossible: first, running backpropagation twice is tedious. Second, we can not easily store a matrix with d^2 entries since d is the number of parameters in the neural language model, which is exceedingly large.

3.3.2 Approximating the Hessian

To cut the computation down to one pass, I exploit a property from theoretical statistics: Namely, that the Hessian of the log-likelihood bears a close resemblance to a quantity known as the Fisher information matrix. This connection allows us to develop a more efficient algorithm that approximates the Hessian with one pass of backpropagation.

I derive this approximation to the Hessian of $\mathcal{L}(\boldsymbol{\vartheta})$ here. First, note that due to the linearity of ∇^2, we have

$$\begin{aligned}\mathbf{H} &= \nabla^2 \mathcal{L}(\boldsymbol{\vartheta}) \\ &= \nabla^2 \left(\sum_{\ell \in \mathcal{T}} \log p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}) + \log p(\boldsymbol{\vartheta}) \right) \\ &= \underbrace{\sum_{\ell \in \mathcal{T}} \nabla^2 \log p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta})}_{\textit{likelihood}} + \underbrace{\nabla^2 \log p(\boldsymbol{\vartheta})}_{\textit{prior}}\end{aligned} \tag{3.9}$$

Note that the integral over languages $\ell \in \mathcal{T}$ is a discrete summation, so we may exchange addends and derivatives such as is required for the proof.

We now discuss each term of Equation (3.9) individually. First, to approximate the likelihood term, I draw on the relation between the Hessian and the Fisher information matrix. A basic fact from information theory Cover and Thomas (2006) gives us that the Fisher information matrix may be written in two equivalent ways:

$$\begin{aligned}&- \mathbb{E}\left[\nabla^2 \log p(\mathcal{D} \mid \boldsymbol{\vartheta})\right] \\ &= \underbrace{\mathbb{E}\left[\nabla \log p(\mathcal{D} \mid \boldsymbol{\vartheta}) \nabla \log p(\mathcal{D} \mid \boldsymbol{\vartheta})^\top\right]}_{\textit{expected Fisher information matrix}}\end{aligned} \tag{3.10}$$

This equality suggests a natural approximation of the expected Fisher information matrix—the *observed* Fisher information matrix

$$-\frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{\mathbf{x}\in\mathcal{D}_\mathcal{T}}\nabla^2\log p(\mathbf{x}\mid\boldsymbol{\vartheta}) \tag{3.11}$$
$$\approx \underbrace{\frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{\mathbf{x}\in\mathcal{D}_\mathcal{T}}\nabla\log p(\mathbf{x}\mid\boldsymbol{\vartheta})\,\nabla\log p(\mathbf{x}\mid\boldsymbol{\vartheta})^\top}_{\textit{observed Fisher information matrix}}$$

which is tight in the limit as $|\mathcal{D}_\mathcal{T}|\to\infty$ due to the law of large numbers. Indeed, when we have a large number of training exemplars, the average of the outer products of the gradients will be a good approximation to the Hessian. However, even this approximation still has d^2 entries, which is far too many to be practical. Thus, I further use a diagonal approximation. I denote the diagonal of the observed Fisher information matrix as the vector $\mathbf{f}\in\mathbb{R}^d$, which I define as

$$\mathbf{f}=\sum_{\ell\in\mathcal{T}}\sum_{\mathbf{x}\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|\cdot|\mathcal{T}|}\Big(\nabla\log p(\mathbf{x}\mid\boldsymbol{\vartheta})\Big)^2 \tag{3.12}$$

Computation of the Hessian of the prior term in Equation (3.9) is more straightforward and does not require approximation. Indeed, in the general case, this is the negative inverse of the covariance matrix, which in our case means

$$\nabla^2\log p(\boldsymbol{\vartheta})=-\frac{1}{\sigma^2}\mathbf{I} \tag{3.13}$$

Summing the (approximate) Hessian of the log-likelihood in Equation (3.12) and the Hessian of the prior in Equation (3.13) yields our approximation to the Hessian of the log-posterior

$$\tilde{\mathbf{H}}=-\mathrm{diag}(\mathbf{f})-\frac{1}{\sigma^2}\mathbf{I} \tag{3.14}$$

3.3.3 MAP Inference

Finally, I harness the posterior $p(\boldsymbol{\vartheta}\mid\mathcal{D}_\mathcal{T})\approx\mathcal{N}(\boldsymbol{\vartheta}^\star,-\tilde{\mathbf{H}}^{-1})$ as the prior over model parameters for training a language model on new, held-out languages via MAP estimation. This is only an approximation to full Bayesian inference, because it does not characterise the entire distribution of the posterior, just the mode (Gelman et al., 2013).

In the zero-shot setting, this boils down to using the mean of the prior $\boldsymbol{\vartheta}^\star$ as network parameters during evaluation. In the few-shot setting, instead, I assume that some data for the target language $\ell \in \mathcal{E}$ is available. Therefore, I maximise the log-likelihood given the target language data plus a regularizer that incarnates the prior, scaled by a factor of λ:

$$\mathcal{L}(\boldsymbol{\vartheta}) = \sum_{\ell \in \mathcal{E}} \log p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}) + \frac{\lambda}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star)^\top \tilde{\mathbf{H}} \, (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\star) \tag{3.15}$$

I denote the the prior $\mathcal{N}(\boldsymbol{\vartheta}^\star, -\tilde{\mathbf{H}}^{-1})$ that features in Equation (3.15) as Univ, as it incorporates universal linguistic knowledge. As a baseline for this objective, I perform MAP inference with an uninformative prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$, which I label Ninf. In the zero-shot setting, this means that the parameters are sampled from the uninformative prior. In the few-shot setting, I maximise

$$\mathcal{L}(\boldsymbol{\vartheta}) = \sum_{\ell \in \mathcal{E}} \log p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}) - \frac{\lambda}{2}||\mathbf{w}||_2^2 \tag{3.16}$$

Note that, owing to this formulation, the uninformed Ninf model does not have access to the posterior of the weights given the data from the training languages.

Moreover, as an additional baseline, I consider a common approach for transfer learning in neural networks (Ruder, 2017), namely 'fine-tuning.' After finding the maximum-likelihood value $\boldsymbol{\vartheta}^\star$ on the training data, this is simply used to initialise the weights before further refining them on the held-out data. I label this method FiTu. In Bayesian terms, this last baseline corresponds to assuming an unnormalisable prior.

3.4 Language Modelling Conditioned on Typological Features

Realistically, the prior over network weights should also be augmented with side information about the general properties of the held-out language to be learned, if such information is available. In fact, linguists have documented such information even for languages without plain digital texts available and stored it in publicly accessible databases (Croft, 2002; Dryer and Haspelmath, 2013). This information usually takes the form of features that express either: i) the formal strategies each language employs to express a

specific semantic or functional construction (Croft et al., 2017). For instance, English expresses the construction of nominal predication with a copula strategy; or ii) the presence or absence of a linguistic category. For instance, English possesses grammatical tense.

The usage of such features to inform neural NLP models is still scarce, partly because the evidence in favour of their effectiveness is mixed (Ponti et al., 2019a, 2018a) and partly because many features are not documented for many languages (Bjerva et al., 2020). In this work, I propose a way to distantly supervise the model with this *side information* effectively. I extend the non-conditional language models outlined in Section 3.3 (BARE) to a series of variants *conditioned* on language-specific properties, inspired by Östling and Tiedemann (2017) and Platanios et al. (2018). A fundamental difference from these previous works, however, is that they learn such properties in an end-to-end fashion from the data in a joint multilingual learning setting. Obviously, this is not feasible for the zero-shot setting and unreliable for the few-shot setting. Rather, I represent languages with their typological feature vector, which I assume readily available both for training and for held-out languages.

Let $\mathbf{t}_\ell \in [0,1]^f$ be a vector of f typological features for language $\ell \in \mathcal{T} \sqcup \mathcal{E}$. The collection of such features for all training languages is denoted by $\mathcal{F}_\mathcal{T}$. I reinterpret the conditional language models within the Bayesian framework by estimating their posterior probability

$$p(\boldsymbol{\vartheta} \mid \mathcal{D}_\mathcal{T}, \mathcal{F}_\mathcal{T}) \propto \prod_{\ell \in \mathcal{T}} p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}, \mathbf{t}_\ell)\, p(\boldsymbol{\vartheta} \mid \mathbf{t}_\ell). \tag{3.17}$$

I now consider two possible methods to estimate $p(\mathcal{D}_\ell \mid \boldsymbol{\vartheta}, \mathbf{t}_\ell)$. For both of them, I first encode the features through a non-linear transformation $f(\mathbf{t}_\ell) = \text{ReLU}(\mathbf{W}\,\mathbf{t}_\ell + \mathbf{b})$, where $\mathbf{W} \in \mathbb{R}^{r \times f}$ and $\mathbf{b} \in \mathbb{R}^r$, $r \ll f$. A first variant, labelled OEST, is based on Östling and Tiedemann (2017). Assuming the standard LSTM architecture where $\mathbf{o}_t$ is the output gate and $\mathbf{c}_t$ is the memory cell, I modify the equation for the hidden state $\mathbf{h}_t$ as follows:

$$\mathbf{h}_t = \Big(o_t \odot \tanh(\mathbf{c}_t)\Big) \oplus f(\mathbf{t}_\ell) \tag{3.18}$$

where $\odot$ stands for the Hadamard product and $\oplus$ for concatenation. In other words, I concatenate the typological features to all the hidden states.

Moreover, I experiment with a second variant where the parameters of the LSTM are generated by a hyper-network (i.e., a simple linear layer with weight $\mathbf{W} \in \mathbb{R}^{|\boldsymbol{\vartheta}| \times r}$) that transforms $f(\mathbf{t}_\ell)$ into $\boldsymbol{\vartheta}$. This approach, labelled PLAT, is inspired by Platanios

et al. (2018), with the additional difference that they generate parameters for an encoder-decoder architecture for neural machine translation, not for a language model.

On the other hand, I do not consider the conditional model proposed by Sutskever et al. (2014), where $f(\mathbf{t}_\ell)$ would be used to initialise the values for $\mathbf{h}_0$ and $\mathbf{c}_0$. During the evaluation, for all time steps t, $\mathbf{h}_t$ and $\mathbf{c}_t$ are never reset on sentence boundaries, so this model would find itself at a disadvantage because it would require either to erase the sequential history cyclically or to lose memory of the typological features over long sequences.

3.5 Experimental Setup

Data The source of text data is the Bible corpus[7] (Christodouloupoulos and Steedman, 2015).[8] I exclude languages that are not written in the Latin script[9] and duplicate languages, resulting in a sub-sample of 77 languages.[10] Since not all texts cover the entire Bible, they vary in size. The text from each language is split into training, development, and evaluation sets with a ratio of 80/10/10%. Moreover, for the MAP inference in the few-shot setting, I randomly sample 100 sentences from each training set.

I obtain the typological feature vectors from URIEL (Littell et al., 2017).[11] I include the features related to 3 levels of linguistic structure, for a total of 245 features: i) syntax, e.g. whether the subject tends to precede the object. These originate from the World Atlas of Language Structures (Dryer and Haspelmath, 2013) and the Syntactic Structures of the World's Languages (Collins and Kayne, 2009); ii) phonology, e.g. whether a language has distinctive tones; iii) phonological inventories, e.g. whether a language possesses the retroflex approximant /ɻ/. Both ii) and iii) were originally collected in PHOIBLE (Moran et al., 2014). Missing values were inferred as a weighted average of the 10 nearest neighbour languages in terms of family, geography, and typology.[12]

[7] http://christos-c.com/bible/

[8] This corpus is arguably representative of the variety of the world's languages: it covers 28 genealogical families, several geographic areas (16 languages from Africa, 23 from Americas, 26 from Asia, 33 from Europe, 1 from Oceania), and endangered or poorly documented languages (39 with less than a million speakers).

[9] The choice of a homogeneous script is necessary because of the assumption that characters are reasonable proxies for underlying phonemes. Multiple scripts could be modelled jointly in the future, provided that a character-to-phoneme mapping is available.

[10] These are identified with their 3-letter ISO 639-3 codes throughout the chapter. Please consult Table B.1 in the Appendix for the full list of language names mapped to ISO 639-3 codes.

[11] http://www.cs.cmu.edu/~dmortens/uriel.html

[12] The heat map of the binary matrix of typological features is shown in Figure B.1 in the Appendix.

	Ninf	Univ			Ninf	Univ			Ninf	Univ	
	Bare	Bare	Oest		Bare	Bare	Oest		Bare	Bare	Oest
acu	8.491	**3.244**	3.472	*fra*	8.587	**4.066**	4.467	*por*	8.491	**3.751**	4.219
afr	8.607	**3.229**	3.995	*gbi*	8.610	**3.823**	3.912	*pot*	8.600	**5.336**	5.359
agr	8.603	**3.779**	3.946	*gla*	8.490	4.179	3.956	*ppk*	8.596	**4.506**	4.599
ake	8.602	**5.753**	6.281	*glv*	8.606	**4.349**	4.612	*quc*	8.605	**4.063**	4.118
alb	8.490	**4.571**	5.017	*hat*	8.594	**4.186**	4.620	*quw*	8.488	**3.560**	4.027
amu	8.610	**4.912**	5.959	*hrv*	8.606	4.050	**3.441**	*rom*	8.603	**3.669**	4.056
bsn	8.591	**5.046**	5.695	*hun*	8.493	**4.836**	5.030	*ron*	8.588	**5.011**	5.690
cak	8.603	**4.068**	4.326	*ind*	8.604	**3.796**	4.311	*shi*	8.601	**5.496**	5.946
ceb	8.488	**3.668**	3.850	*isl*	8.596	**5.039**	5.629	*slk*	8.491	**4.304**	4.512
ces	8.600	**4.369**	4.461	*ita*	8.605	4.023	**3.752**	*slv*	8.604	**3.661**	4.106
cha	8.594	4.366	**4.353**	*jak*	8.488	**4.051**	4.793	*sna*	8.596	**4.146**	4.283
chq	8.598	**6.940**	7.623	*jiv*	8.601	**3.866**	4.039	*som*	8.614	**4.159**	4.470
cjp	8.494	**4.600**	4.985	*kab*	8.596	**4.659**	5.400	*spa*	8.489	**3.645**	4.020
cni	8.604	**3.740**	4.651	*kbh*	8.607	**4.663**	4.950	*srp*	8.604	**3.414**	3.437
dan	8.593	**3.471**	4.599	*kek*	8.491	**4.666**	4.944	*ssw*	8.593	4.064	**3.780**
deu	8.599	**4.102**	4.214	*lat*	8.601	**3.703**	4.093	*swe*	8.605	4.210	**3.892**
dik	8.490	**4.447**	4.533	*lav*	8.588	**5.415**	6.130	*tgl*	8.487	**3.639**	3.878
dje	8.603	**3.725**	3.996	*lit*	8.602	**4.794**	4.853	*tmh*	8.602	4.830	**4.711**
djk	8.592	**3.663**	3.874	*mam*	8.488	**4.292**	5.076	*tur*	8.592	**5.574**	5.935
dop	8.609	**5.950**	7.351	*mri*	8.606	**3.440**	4.074	*usp*	8.604	**4.127**	4.337
eng	8.488	**3.816**	4.028	*nhg*	8.588	**4.323**	4.450	*vie*	8.490	**7.137**	7.484
epo	8.605	**3.818**	4.116	*nld*	8.601	**3.851**	4.326	*wal*	8.605	**4.027**	4.585
est	8.606	**6.807**	8.261	*nor*	8.492	**3.174**	3.902	*wol*	8.607	**4.290**	4.420
eus	8.605	**4.118**	4.321	*pck*	8.603	**4.053**	4.233	*xho*	8.602	**4.171**	4.276
ewe	8.490	**5.049**	5.497	*plt*	8.603	**4.364**	4.648	*zul*	8.488	**3.218**	4.109
fin	8.604	**4.308**	4.338	*pol*	8.601	**5.158**	5.556	ALL	8.572	**4.343**	4.691

Table 3.1 BPC scores (lower is better) for the ZERO-SHOT learning setting, with the uninformed prior (NINF) and the universal prior (UNIV): see §3.2 for the descriptions of the priors. Colors define the split in which each language (rows) has been held out.

	Bare	Oest		Bare	Oest		Bare	Oest		Bare	Oest
acu	1.413	**1.308**	*eng*	1.355	**1.350**	*kek*	**1.131**	1.133	*slk*	1.844	**1.754**
afr	1.471	**1.457**	*epo*	1.471	**1.450**	*lat*	1.792	**1.758**	*slv*	1.848	**1.793**
agr	1.701	**1.581**	*est*	0.333	**0.150**	*lav*	2.146	**1.931**	*sna*	1.489	**1.457**
ake	1.453	**1.377**	*eus*	1.763	**1.635**	*lit*	1.895	**1.833**	*som*	1.477	**1.468**
alb	1.590	**1.552**	*ewe*	2.084	**1.944**	*mam*	1.654	**1.548**	*spa*	1.559	**1.525**
amu	1.402	**1.340**	*fin*	1.716	**1.680**	*mri*	1.342	**1.330**	*srp*	1.832	**1.756**
bsn	1.232	**1.172**	*fra*	1.465	**1.432**	*nhg*	1.302	**1.238**	*ssw*	1.890	**1.697**
cak	1.281	**1.221**	*gbi*	1.398	**1.331**	*nld*	1.621	**1.601**	*swe*	1.619	**1.595**
ceb	1.193	**1.185**	*gla*	3.403	**1.839**	*nor*	1.623	**1.590**	*tgl*	1.221	**1.210**
ces	1.872	**1.795**	*glv*	1.932	**1.644**	*pck*	1.731	**1.711**	*tmh*	2.786	**2.301**
cha	1.934	**1.790**	*hat*	1.480	**1.454**	*plt*	1.296	**1.286**	*tur*	1.801	**1.773**
chq	1.265	**1.220**	*hrv*	2.059	**1.974**	*pol*	1.743	**1.698**	*usp*	1.290	**1.214**
cjp	1.706	**1.565**	*hun*	1.887	**1.847**	*por*	1.586	**1.552**	*vie*	1.648	**1.637**
cni	1.348	**1.290**	*ind*	1.356	**1.336**	*pot*	2.484	**2.144**	*wal*	1.561	**1.457**
dan	1.727	**1.693**	*isl*	1.845	**1.808**	*ppk*	1.538	**1.439**	*wol*	2.053	**1.890**
deu	1.532	**1.512**	*ita*	1.615	**1.583**	*quc*	1.393	**1.291**	*xho*	1.680	**1.634**
dik	1.979	**1.835**	*jak*	1.415	**1.322**	*quw*	1.498	**1.418**	*zul*	1.880	**1.620**
dje	1.570	**1.550**	*jiv*	1.705	**1.572**	*rom*	1.706	**1.587**	ALL	1.652	**1.550**
djk	1.515	**1.435**	*kab*	1.955	**1.791**	*ron*	1.572	**1.537**			
dop	1.810	**1.676**	*kbh*	1.436	**1.371**	*shi*	2.057	**1.903**			

Table 3.2 BPC results (lower is better) for the JOINT learning setting, with the uninformed NINF prior. These results constitute the ceiling performance for language transfer models.

	Ninf	FiTu	Univ			Ninf	FiTu	Univ	
	Bare	Oest	Bare	Oest		Bare	Oest	Bare	Oest
acu	4.203	**2.117**	2.551	2.136	*kbh*	4.644	2.362	2.434	**2.288**
afr	4.423	3.620	3.042	**2.773**	*kek*	4.613	2.809	3.015	**2.714**
agr	4.268	3.282	3.403	**2.457**	*lat*	4.239	4.342	3.416	**3.202**
ake	4.318	2.168	2.238	**2.180**	*lav*	4.765	**2.867**	3.842	2.917
alb	4.544	3.186	3.302	**3.084**	*lit*	4.769	3.752	**3.592**	3.668
amu	4.486	2.820	3.948	**2.080**	*mam*	4.525	**2.274**	2.873	2.363
bsn	4.546	1.861	2.678	**1.850**	*mri*	3.795	3.482	3.010	**2.459**
cak	4.426	1.994	2.053	**1.956**	*nhg*	4.373	2.004	2.480	**1.965**
ceb	4.084	2.562	2.595	**2.470**	*nld*	4.469	3.008	2.908	**2.903**
ces	4.984	4.651	4.190	**3.680**	*nor*	4.453	3.152	**2.954**	3.054
cha	4.329	2.546	2.899	**2.525**	*pck*	4.246	4.011	3.532	**3.030**
chq	4.941	**1.948**	2.078	1.963	*plt*	4.201	2.532	2.742	**2.490**
cjp	4.424	**2.389**	2.880	2.393	*pol*	4.853	3.852	**3.620**	3.788
cni	4.185	2.797	3.018	**1.982**	*por*	4.446	3.231	3.198	**3.098**
dan	4.719	3.211	**3.127**	3.180	*pot*	4.299	3.773	3.944	**2.763**
deu	4.589	3.103	3.007	**2.953**	*ppk*	4.439	**2.220**	2.736	2.236
dik	4.380	**2.640**	3.020	2.667	*quc*	4.538	2.154	2.242	**2.108**
dje	4.382	3.815	3.398	**2.898**	*quw*	4.223	2.196	2.547	**2.158**
djk	4.130	**2.064**	2.446	2.085	*rom*	4.378	3.121	3.257	**2.455**
dop	4.508	2.506	2.562	**2.448**	*ron*	4.579	3.273	3.734	**3.216**
eng	4.436	2.808	2.913	**2.719**	*shi*	4.509	**2.963**	3.092	2.970
epo	4.469	3.609	3.511	**2.825**	*slk*	4.873	3.722	3.812	**3.631**
est	3.618	**1.952**	2.487	1.962	*slv*	4.633	4.630	3.527	**3.501**
eus	4.354	2.628	2.705	**2.567**	*sna*	4.455	2.910	3.114	**2.870**
ewe	4.590	2.806	3.336	**2.786**	*som*	4.257	3.048	**2.908**	2.934
fin	4.385	4.339	3.830	**3.312**	*spa*	4.507	3.223	3.149	**3.090**
fra	4.551	3.086	3.276	**2.981**	*srp*	4.561	4.467	**3.367**	3.380
gbi	4.250	2.138	2.170	**2.054**	*ssw*	4.370	2.611	2.924	**2.570**
gla	4.159	**2.377**	2.835	2.395	*swe*	4.657	3.266	3.184	**3.177**
glv	4.346	3.523	3.702	**2.644**	*tgl*	4.060	2.546	2.592	**2.436**
hat	4.468	2.929	3.048	**2.849**	*tmh*	4.618	4.087	4.218	**3.125**
hrv	4.615	3.845	3.608	**3.588**	*tur*	4.846	**3.509**	4.282	3.552
hun	4.806	3.589	3.709	**3.522**	*usp*	4.529	2.114	2.189	**2.073**
ind	4.377	3.317	3.258	**2.420**	*vie*	5.185	3.018	3.751	**3.015**
isl	4.744	3.174	3.703	**3.101**	*wal*	4.398	2.986	3.623	**2.278**
ita	4.370	3.384	3.196	**3.178**	*wol*	4.621	2.898	2.968	**2.826**
jak	4.532	2.113	2.650	**2.126**	*xho*	4.561	3.415	**3.208**	3.289
jiv	4.338	3.413	3.475	**2.504**	*zul*	4.564	2.625	2.866	**2.622**
kab	4.649	**2.783**	3.574	2.800	ALL	4.467	3.007	3.120	**2.731**

Table 3.3 BPC scores for the FEW-SHOT learning setting, with NINF, FITU and UNIV priors. Colors define the split in which each language (rows) has been held out.

Language Model I implement the LSTM following the best practices and hyper-parameter settings indicated for language modelling by Merity et al. (2017, 2018). In particular, I tie input and output embeddings and optimise the weights with Adam (Kingma and Ba, 2015) and a non-monotonically decayed learning rate: its value is initialised as 10^{-4} and decreases by a factor of 10 every 1/3rd of the total epochs. The maximum number of epochs amounts to 6 for training, with early stopping based on development set performance, and the maximum number of epochs is 25 for few-shot learning.

For each training iteration, I sample a language proportionally to the amount of its data: $p(\ell) \propto |\mathcal{D}_\ell|$, in order not to exhaust examples from resource-lean languages in the early phase of training. Then, I sample without replacement from $\mathcal{D}_\ell$ a mini-batch of 128 sequences with a variable maximum sequence length.[13] This length is sampled from a distribution $m \sim \mathcal{N}(\mu = 125, \sigma = 5)$.[14] Each epoch comes to an end when all the data sequences have been sampled.

I apply several techniques of dropout for regularisation, including variational dropout (Gal and Ghahramani, 2016b), which applies an identical mask to all time steps, with $p = 0.1$ for character embeddings and intermediate hidden states, and $p = 0.4$ for the output hidden states. DropConnect (Wan et al., 2013) is applied to the model parameters U of the first hidden layer with $p = 0.2$.

Following Merity et al. (2017), the underlying language model architecture consists of 3 hidden layers with 1,840 hidden units each. The dimensionality of the character embeddings is 400. For conditional language models, the dimensionality of $f(\mathbf{t})$ is set to 115 with the OEST method based on concatenation (Östling and Tiedemann, 2017), and 4 (due to memory limitations) in the PLAT method based on meta-networks (Platanios et al., 2018). For the regulariser in Equation (3.15), I performed grid search over the hyperparameter λ: I finally select a value of 10^5 for UNIV and 10^{-5} for NINF.

Regimes of Data Paucity I explore the following regimes of data paucity for the held-out languages:

• ZERO-SHOT transfer setting: I split the sample of 77 languages into 4 subsets. The languages in each subset are held out in turn, and I use their test set for evaluation.[15] For each subset, I further randomly choose 5 languages whose development set is used

[13] This avoids creating insurmountable boundaries to back-propagation through time (Tallec and Ollivier, 2017).

[14] The learning rate is therefore scaled by $\frac{m}{\mu} \times \frac{|\mathcal{D}_\mathcal{T}|}{L \cdot |\mathcal{D}_\ell|}$, where L is the total number of languages, on account of the variability in sequence length and in training data per language.

[15] Holding out each language individually would not increase the sample of training languages significantly, while inflating the number of experimental runs needed.

for validation. The training set of the rest of the languages is used to estimate a prior over network parameters via the Laplace approximation.

• FEW-SHOT transfer setting: on top of the zero-shot setting, I use the prior to perform MAP inference over a small sample (100 sentences) from the training set of each held-out language.

• JOINT multilingual setting: $\mathcal{T} = \mathcal{E}$ such that the full training set for all 77 languages is observed, without any held-out language. This works as a ceiling for the expected performance of language transfer models.

3.6 Results and Analysis

The results for our experiments are grouped in Table 3.1 for the ZERO-SHOT regime, Table 3.3 for the FEW-SHOT regime, and in Table 3.2 for the JOINT multilingual regime. The scores represent Bits Per Character (BPC) (Graves, 2013): this metric is simply defined as the average negative log-likelihood of test data divided by $\log 2$. I compare the results along the following dimensions:

Informativeness of Prior The main result is that the UNIV prior consistently outperforms both baselines by a large margin in both ZERO-SHOT and FEW-SHOT settings. The superiority over the NINF prior suggests that transfer is possible in the first place, since cross-lingual tendencies give a faithful picture of what to expect in new languages. The *lowest* BPC reductions are observed for languages like Vietnamese (15.94% error reduction) or Highland Chinantec (19.28%) where character distributions are unmatched in other languages. As for the FITU prior, the universal prior reduces the average BPC error from 3.007 to 2.731 (in the OEST conditional model). This indicates a second important conclusion: that correct uncertainty information in the prior makes learning more accurate and sample-efficient. Hence, the proposed method holds promise to undermine fine-tuning as the most reliable mechanism of cross-lingual transfer (Peters et al., 2019). Thirdly, note that the ZERO-SHOT UNIV models are on a par or better than even the FEW-SHOT NINF models. In other words, the most helpful supervision comes from a universal prior rather than from a small in-language sample of sentences. All these results demonstrate that the UNIV prior is truly imbued with universal linguistic knowledge that facilitates learning of previously unseen languages.

Conditioning on Typological Information Another fascinating result regards the fact that conditioning language models on typological features yields opposite effects in

the Zero-shot setting compared to the Few-shot and Joint multilingual settings. By comparing the Bare and Oest models' columns in Table 3.1, the non-conditional baseline Bare is superior for 71 / 77 languages (the exceptions being Chamorro, Croatian, Italian, Swazi, Swedish, and Tuareg). On the other hand, the same columns in Table 3.3 and Table 3.2 reveal an opposite pattern: Oest outperforms the Bare baseline in 70 / 77 languages. Finally, Oest surpasses the Bare baseline in the Joint setting for 76 / 77 languages (save Q'eqchi').

I also took into consideration an alternative conditioning method, namely Plat. For clarity's sake, I exclude this batch of results from Table 3.1 and Table 3.3, as this method proves to be consistently worse than Oest. In fact, the average BPC of Plat amounts to 5.479 in the Zero-shot setting and 3.251 in the Few-shot setting. These scores have to be compared with 4.691 and 2.731 for Oest, respectively.

A possible explanation behind the mixed evidence on the success of typological features, on the one hand, points to some intrinsic flaws of typological databases. Ponti et al. (2019a) have shown how i) the feature granularity may be too coarse to liaise with data-driven, exemplar-based probabilistic models; ii) the limited coverage of features results in noise introduced by the inferred missing values; and iii) database information is restricted to the majority strategy within a language and overshadows language-internal variation, hence hindering models from learning less likely but plausible patterns (Sproat, 2016).

Arguably, another cause of the failure of the typology-informed model in the zero-shot setting may be connected to how the model exploits the typological features. Possibly, the model uses features as indices to memorise parameter configurations, irrespective of their original typological interpretation. Thus, at least a few examples are necessary to create the correct mapping between an index and its corresponding parameters. However, the features are still useful as they softly tie the parameters of languages with similar properties. As a consequence of both these reasons, language models seem to be damaged by typological features in absence of data, whereas they find a way to follow their lead when at least a small sample of sentences is available in the Few-shot setting.

Data Paucity Different regimes of data paucity display uneven levels of performance. The best models for each setting (Zero-shot Univ Bare, Few-shot Univ Oest, and Joint Oest) reveal large gaps between their average scores. Hence, in-language supervision remains unsubstitutable, as transferred language models still lag behind their resource-rich equivalents.

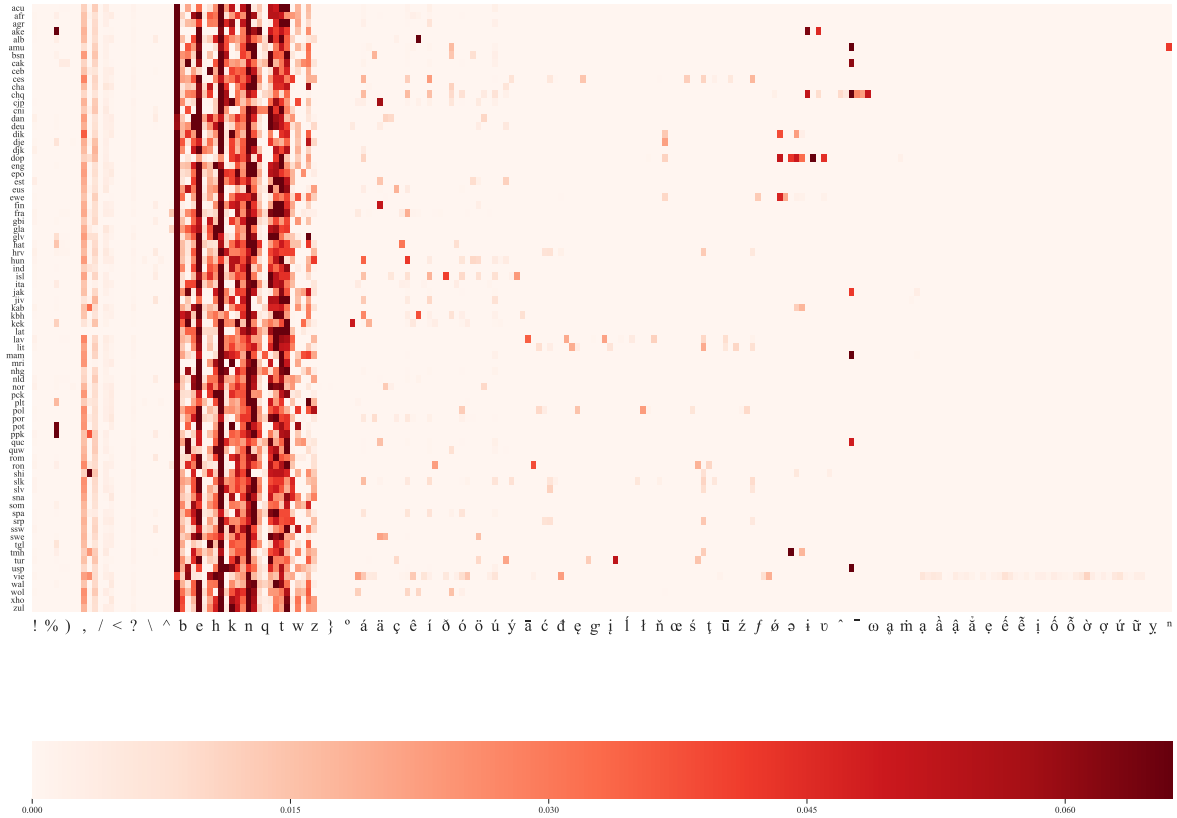


Figure 3.1 Unigram character distribution (x-axis) per language (y-axis). Note how some rows stand out as outliers.

Character Distribution Even within the same setting, BPC scores vary enormously across languages in both the ZERO-SHOT and FEW-SHOT settings, which requires an explanation. Similarly to Gerz et al. (2018a,b), I run a correlation analysis between language modelling performance and basic statistics of the data. In particular, I first create a vector of unigram character frequencies for each language, resulting in the matrix shown in Figure 3.1. Then I estimate the average cosine distance between the vector of a held-out language and each training language. This returns a measure of the 'exoticness' of a language's character distribution.

Pearson's correlation between such cosine distance and the perplexity of UNIV BARE in each language reveals a strong correlation coefficient $\rho = 0.53$ and a statistical significance of $p < 10^{-6}$ in the ZERO-SHOT setting. On the other hand, such correlation is absent ($\rho = -0.13$) and insignificant $p > 0.2$ in the FEW-SHOT setting. In other

words, language models cease to depend on the source unigram character distribution and quickly adapt to the target one after only few examples.

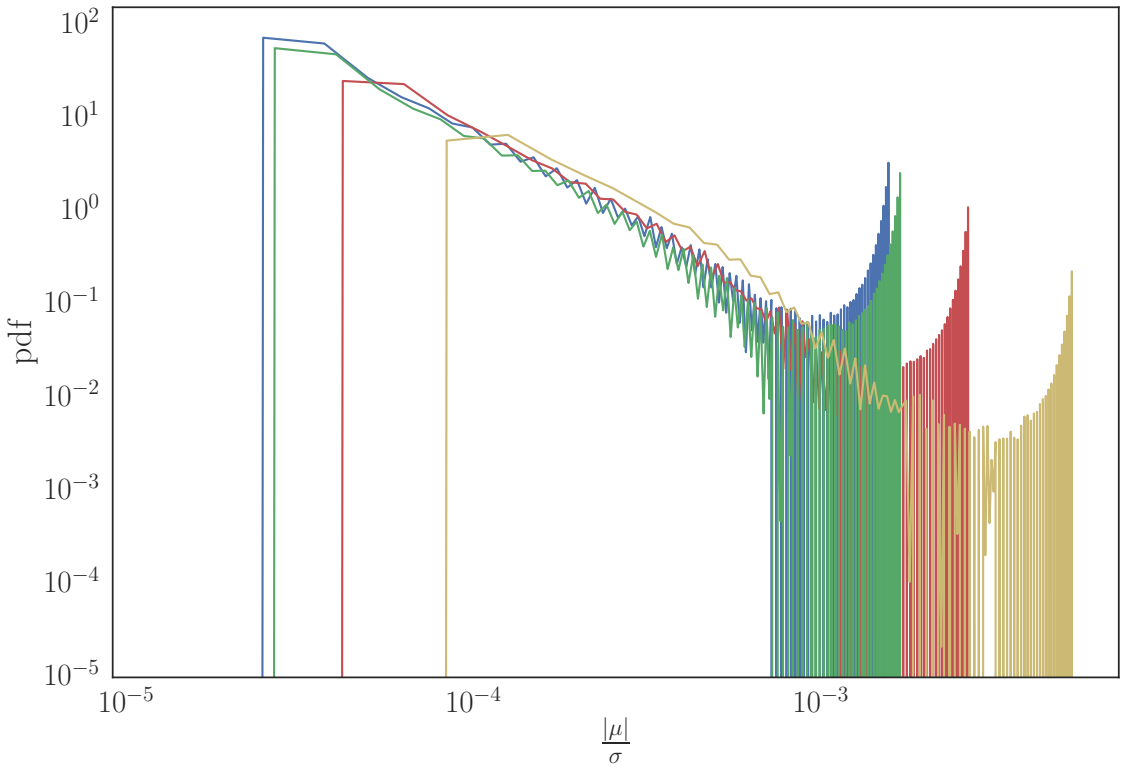


Figure 3.2 Probability density function of the signal-to-noise ratio for each parameter of the learned posteriors in the Univ Bare language models on splits 1 (blue), 2 (red), 3 (green), 4 (gold). The plot is in log-log scale.

Probing of Learned Posteriors Finally, it remains to establish which sort of knowledge is embedded in the universal prior. How to probe a probability distribution over weights in the non-conditional Univ Bare language model? First, I study the signal-to-noise ratio of each parameter $\boldsymbol{\vartheta}_i$, computed as $\frac{|\mu_i|}{\sigma_i}$, in each of the 4 splits. Intuitively, this metric quantifies the 'informativeness' of each parameter, which is proportional to both the absolute value of the mean and the certainty of the estimate. The probability density function of the signal-to-noise ratio is shown in Figure 3.2. From this plot, it emerges that the uncertainty is generally under-estimated (small σ_i denominators yield high values). Most crucially, the signal-to-noise values concentrate on the left of the spectrum, which means they will incur in any penalty while changing, based on Equation (3.15); on the other hand, there is a bulk of highly informative parameters on the right of the spectrum

that are very likely to remain adamant, thus preventing catastrophic forgetting. All splits display such a pattern, although somewhat shifted.

Second, to study the effect of conditioning the universal prior on typological features, I generate random sequences of 25 characters from the learned prior in each language. The first character is chosen uniformly at random, and the subsequent ones are sampled from the distribution given by Equation (3.1) with a temperature of 1. The resulting texts are shown in Table 3.4. Although this would warrant a more thorough and systematic analysis, from a cursory view it is evident of the sequences abide with universal phonological patterns, e.g. favouring vowels as syllabic nuclei and ordering consonants based on sonority hierarchy. Moreover, the language-specific information clearly steers predicted sequences towards the correct inventory of characters, as demonstrated by Vietnamese (VIE) and Lukpa (DOP) in Table 3.4.

Sources of Uncertainty A crucial underlying assumption of this experiment is that a more expressive prior, which takes uncertainty into account, better reflects the cross-linguistic variability. Crucially, it must be noted that the variation is inherent to the quantity being modelled, as its distribution spans different parameter configurations, which are language-specific. Therefore, it would persist even in the limit of infinite data, where sample-size effects vanish. Clearly, however, every experiment operates in a scenario with limited data: hence, both sources of uncertainty are present in our model. Within Bayesian theory, they are conflated and treated as one and the same. Therefore, it is not possible to disentangle these two components in the learned prior. If infinite data were available, the distribution would better reflect high-probability areas and thus better guide the parameters towards those during fast few-shot learning.

3.7 Related Work

LSTM architectures have been probed for an inductive bias in capturing syntactic dependencies (Linzen et al., 2016) and grammaticality judgements (Marvin and Linzen, 2018; Warstadt et al., 2019). Ravfogel et al. (2019) have extended the scope of this analysis to typologically different languages through *synthetic* variations of English. In this work, I modelled the inductive bias explicitly by constructing a prior over the space of neural network parameters.

Few-shot word-level language modelling for truly under-resourced languages such as Yongning Na has been investigated by Adams et al. (2017) with the aid of a bilingual lexicon. Vinyals et al. (2016) and Munkhdalai and Trischler (2018) proposed novel

LIT	*javen šuksyr sun siriai tes pije nuks*	SHI	*ereswrin an daɣtartnaas ni mad yanó*
NOR	*s hech far binje alrn bre a ver e hior*	JAK	*fi pelo ayok musam nejaz jih tewat ushi*
KEK	*sx er taj chan linam laj âtebke naque*	SWE	*ssiar řades perdeshen heklui tart si a*
JIV	*da tum suuam sɪtas nekkin una tekaru ni*	DIK	*e wɛn ke nuŋ ni piyitia de run ye e ke*
DJE	*a ciya toi milkak mo to yen nga suci*	EWE	*å mula pe ose le ake mente amesa ke kul*
SLK	*o je to temokoé lostave sa jesé gukli*	ALB	*I kur je ki thet je ji tin nuk t tho*
CES	*e je jek jem neuteŋ rekssýj jazá níb ws*	CNI	*u pen mireshisinoe airitcsa ateani yi*
POR	*uč somo ai jegparase saves e iper to*	POT	*neta ynimka nekin linaayi meu carií a*
SPA	*esquár y lues dusme allis nencec adi*	ZUL	*ŏnakan kuná bencro krileke konusti k*
GLV	*ayr shżi ayn ai sephson a gil or geee*	QUW	*ai chimira kachisinyra poi apre asyu*
POL	*eteni na hidi cẽho oż swchj jeci i cil*	AGR	*ji ica ama kujaa muri wajetar aumam hu*
QUC	*ûs xe cä wija ro pio kin cbi' ij jejac*	DOP	*btɛlɔ ɪ telaɣa kɔ nɛɪ zûɣɪ nɛkə pɔ*
WAL	*banjake la dos que benthi shivegina*	EUS	*cerer nagcermac istirinun qatserite*
XHO	*ukayla azigeecoa kosubentisiili jen maky*	HUN	*elyet a bukot aky azraá ot mu háláj y*
SOM	*ao kun adku i sir jija i befey yadui*	GLA	*o e kere hhó sho dhöìr te ilailui a tu a*
TGL	*ikugy peo asha atan kao amai kain ak a*	PCK	*u gihiha ki mi dhia mea la hen a puh ih*
CJP	*pae yei aje kin trheka pän awawa ri s*	AFR	*mal hoor in e sheei wer var buerkeas en*
ACU	*animmhi mustatur tukaw aants aastasai a*	USP	*okan mi ykis ris rajajkujij taka ja*
FIN	*i koin suu meit ja ii soi tetot jasw*	IND	*t berka duhah menkad kemia ukus keri ya*
MRI	*oki ka benoka ai ki kimanka pikaka ko*	ROM	*hal kus seke nukertia dehe neshes hos n*
SLV	*čičvim koko si neče pau ku meta noj ne*	TMH	*ǝrofm sibarn awigtir ϵli d usi leped*
HRV	*ca ka te zet jon jem nezin isak ve u*	ITA	*tri cordia io si si conse de namni nel*
EPO	*j li inij keris ec xom el e sepon kaj*	SRP	*e se a nil do zasom kuz je sefe nij hoč*
AMU	*ḿibinya na ñero melee cano' ndo' cy'oc*	NLD	*e suet en de semeshord ak abaido zin*
KBH	*ẍe aquangmomnaynangmuacha tojam*	LAT	*ifte quissi fetam remnas emens in timnex*
CEB	*abithon kayay isa atoug giraban sula*	MAM	*í la ŋil a cheh tjea nut tej quxen kaj*
GBI	*fuma ome pani de imoako kema kaye ntul*	VIE	*hả kì đãi bi ât ni γì sa hiổ vū r*
ENG	*g ban urse auth ahen ant msesher at nhe*		
ISL	*j noka nie leli maken ti aide ni itsim a*	EST	*inam acha dius dempegun geben parug j*
SNA	*xe yare ske tengker ci bendar nu derbe*	CHA	*ê duka ka kina kia nextis ne aka nisa*
RON	*ma awa nasil ko khe ni koy koj tikis t*	FRA	*dis assan in man usia issokoj mulel e me*
KAB	*je cana ka casa chomdis mear de ber h*	DJK	*okrana anginar matom iliantarinta a non*
NHG	*chun neyal den ma kashtaka asa as riste*	LAV	*ilu kagsa eriri isi paj ewri bus os*
DAN	*dnepse aa aye sas ningli inas giksaj abe*	BSN	*as juhma yainawa nusa wali apai basti*
PPK	*ios yena mona kemewascoj ni ne maa*	HAT	*a kuneati ua veskos oramaj meseqen ye k*
SSW	*nta yoti gesi kela nii ikasgaber ni tus*	TUR	*che a shachmo ẽspi meng rinnaj e ish em*
WOL	*alen kokpan fed man benu pei ei kestam*	AKE	*n jes silem semmo caja arka wagtoa doo*
DEU	*ke giko si obi rer nin eber tun ke ele*	CHQ	*shas nej neysakun kina alistad mesabe*
CAK	*tej je awem titoj lunik c'u chis m ni*	PLT	*ʋwi meyak me imai anet alavis edte kin*

Table 3.4 Randomly generated text on observed languages (top) and held-out languages (bottom) in the 4th split.

architectures (Matching Networks and LSTMs augmented with Hebbian Fast Weights, respectively) for rapid associative learning in English, and evaluated them in few-shot cloze tests. In this respect, this work is novel in pushing the problem to its most complex formulation, zero-shot inference, in taking into account the largest sample of languages for language modelling to date, and recasting cross-lingual neural transfer into a Bayesian framework.

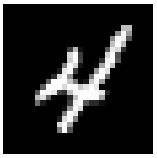
In addition to the set of approaches considered in our work, there are other alternatives to conditional language modelling. Kalchbrenner and Blunsom (2013) used encoded features as additional biases in recurrent layers. Kiros et al. (2014) put forth a log-bilinear model that allows for a 'multiplicative interaction' between hidden representations and input features (such as images). With a similar device, but a different gating method, Tsvetkov et al. (2016) trained a phoneme-level joint multilingual model of words conditioned on typological features from Moran et al. (2014).

The use of the Laplace method for neural continuous learning has been proposed by Kirkpatrick et al. (2017), inspired by synaptic consolidation in neuroscience to avoid catastrophic forgetting. Kochurov et al. (2018) tackled the same problem by approximating the posterior probabilities through stochastic variational inference. Ritter et al. (2018) substitute diagonal Laplace approximation with a Kronecker factored method. Finally, the regulariser proposed by Duong et al. (2015) for cross-lingual dependency parsing can be interpreted as a prior for MAP estimation where the covariance is an identity matrix.

3.8 Conclusions

In this chapter, I proposed a Bayesian approach to cross-lingual language modelling transfer. I created a universal prior over neural network weights that is capable of generalising well to new languages riddled by data paucity, by Laplace-approximating the posterior of the weights given a sample of training languages. Based on the results of character-level language modelling on a sample of 77 languages, I demonstrated the superiority of the universal prior over uninformative priors and uniform priors (i.e., the widespread 'fine-tuning' approach) in both zero-shot and few-shot settings. Moreover, I showed that adding language-specific side information drawn from typological databases to the universal prior further increases the levels of performance in the few-shot regime, although the evidence is mixed in the zero-shot regime. While I also showed that language transfer still lags behind supervised learning when abundant in-language data are available, this work joins current efforts towards bridging this gap in the future. In the

next chapter, I will extend this idea further: first, showing how the idea of constructing an inductive bias can be extended to neural architectures. Second, by demonstrating that this approach can be equally successful in semantic tasks, in addition to structural tasks like language modelling.

Chapter 

A Prior over Architectures for Language Understanding

4.1 Introduction

Constructing a prior exclusively over weight parameters, as endeavoured in Chapter 3, is insufficient to endow artificial neural networks with the correct inductive bias towards natural languages. In fact, as argued in Section 2.2.2, feed-forward functions are fully defined not just in terms of neural weights, but also in terms of what is usually treated as fixed hyper-parameters: layer depth and width, as well as the choice of non-linear activations. In this chapter, I aim at facilitating sample-efficient natural language processing by jointly constructing a prior over weight *and architecture* parameters.

Differentiable Neural Architecture Search (NAS) allows for performing inference on both sets of parameters in an end-to-end fashion through back-propagation (Elsken et al., 2019). This process, however, requires expensive second-order differentiation and two separate stages to optimise first the architecture and then the weights (Liu et al., 2019). Recent developments in the domain of vision, however, solved both issues by modelling the architecture as a variable with categorical distribution (Maddison et al., 2017) and optimising it simultaneously with weight parameters through regular gradient descent (Xie et al., 2019).

In this chapter, I adjust these ideas to natural language processing tasks and architectures, and in particular to state-of-the-art encoders pre-trained on multilingual language modelling (Conneau et al., 2020). Moreover, I reinterpret NAS as empirical Bayes (see Section 2.3.2), as it implicitly defines a hierarchical Bayesian model where a point

estimate of the architecture parameters is inferred through a truncated approximation of maximum-likelihood weights. Under this interpretation, NAS becomes amenable of full Bayesian inference, generalising the original formulation of bi-level optimisation (Liu et al., 2019). For instance, contrary to NAS, I model the dependence of the parameters $\boldsymbol{\vartheta}$ from the architecture $\boldsymbol{\alpha}$ by parameterising the conditional probability $p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha})$ via a hyper-network (see Section 4.3). Similarly to Chapter 3, after that a posterior distribution is obtained from seen language data, I subsequently leverage it for zero-shot and few-shot learning in held-out languages.

In order to evaluate such approach, the ideal dataset must meet a series of desiderata: i) it must be typologically diverse enough to 'stress test' the robustness of cross-lingual transfer towards languages displaying a variety of linguistic features; in other words, it must be specifically tailored to reflect a realistic low-resource setting; ii) it must represent a task related to natural language understanding, in order to demonstrate that the benefit of an inductive bias transcends sequence prediction tasks related to structural knowledge (explored in Chapter 3). In this case, the inductive bias should facilitate reasoning and reflect high-level, abstract knowledge.

Unfortunately, datasets for natural language understanding that are truly typologically diverse are rare, with the notable exception of TyDiQA (Clark et al., 2020). However, passage-based question answering mostly relies on *explicit* information in the text; therefore, the potential for transfer is limited. For these reasons, in this chapter I also detail the creation of a novel dataset for causal commonsense reasoning, the Cross-lingual Choice of Plausible Alternatives (XCOPA; Ponti et al., 2020). XCOPA covers 12 languages in total, including radically under-resourced languages such as Haitian Creole and Southern Quechua. Moreover, solving this benchmark requires to complement explicit textual information with *implicit* knowledge of typical causes and outcomes of real-world situations. Thus, this dataset satisfies all the desiderata to evaluate the Bayesian NAS model proposed in this chapter.

Based on the experimental results, I validate the proposition that positing a prior over parameters and architectures in fact yields gains over state-of-the-art uninformed baselines, which rely on pre-training and fine-tuning. Moreover, these findings demonstrate that the prior constructed in this chapter can in fact enshrine knowledge that helps to enhance the model's reasoning capabilities.

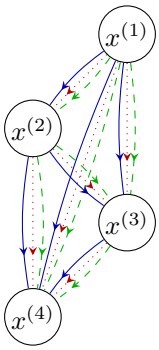


Figure 4.1 Structure of a cell with 4 nodes. Arrows with different colour / style denote different operations.

4.2 Differentiable Neural Architecture Search

Inference over neural architectures was first formulated in a fully differentiable fashion in the seminal work of Liu et al. (2019). In this framework, the search space is that of a feed-forward network *cell* κ whose architecture is parameterized by $\boldsymbol{\alpha}$. Supposing that the hidden state of the network lives in $\mathbb{R}^d$, the cell is a deterministic function that takes as input encoded token representations and the weight parameters $[\mathbf{e}(\mathbf{x}_0, \ldots, \mathbf{x}_n), \boldsymbol{\vartheta}]$ and outputs the hidden representations $(\mathbf{h}_0^{(k+1)}, \ldots, \mathbf{h}_n^{(k+1)})$.

Within a cell, each encoded token input undergoes a series of transformations constituting a directed acyclic graph (DAG) with k (topologically) ordered nodes, each an intermediate representation $(\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(k)})$ where $\mathbf{h}^{(1)} \triangleq \mathbf{e}(\mathbf{x})$. The directed edges of the graph connect each such node with all subsequent nodes, for a total of $\binom{k}{2}$ edges, the $(k-1)^{\text{th}}$ triangular number. Each edge $k_i \to k_j$ corresponds to a transformation characterized by an affine mapping parameterized as $\boldsymbol{\vartheta}^{(i,j)} \triangleq [\mathbf{W}^{(i,j)}, \mathbf{b}^{(i,j)}]$ followed by an operation $o^{(i,j)}$ (e.g. a non-linear activation). The intermediate representation is obtained by reducing all incoming edges through summation:

$$\mathbf{h}^{(i)} = \sum_{j=1}^{i-1} o^{(j,i)} \left(\mathbf{W}^{(j,i)} \mathbf{h}^{(j)} + \mathbf{b}^{(j,i)} \right) \tag{4.1}$$

In particular, I consider the set of operations $O = \{\text{sigmoid}, \tanh, \text{ReLU}, f_I, f_0\}$, where f_I is the identity function and $f_0 \triangleq f : \mathcal{R} \to 0$. This set defines a discrete variable with a categorical distribution. An example of the structure of a cell with 4 nodes is summarised in Figure 4.1. In turn, the cell output representations are the mean of all intermediate representations:

$$\mathbf{h}^{(k+1)} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{h}^{(i)}. \tag{4.2}$$

How to seek simultaneously the optimal weight parameters *and* operations given the data, through back-propagation, in this setting?

DARTS The solution proposed by Liu et al. (2019) is relaxing the distribution over operations into a continuous variable. Rather than an affine transformation followed by a single operation $o^{(i,j)}$ then, an edge becomes a weighted sum of all operations, where the mixing weights are parameterized by $\boldsymbol{\alpha} \in \mathbb{R}^{|O|\times k}$. A softmax ensures that the mixing weights add up to 1. As a consequence, each edge becomes:

$$\mathbf{h}^{(i)} = \sum_{o\in O}\frac{\exp\alpha_o^{(i,j)}}{\sum_{o'\in O}\exp\alpha_{o'}^{(i,j)}}o(\cdot) \tag{4.3}$$

Liu et al. (2019) jointly optimise $\boldsymbol{\vartheta}$ and $\boldsymbol{\alpha}$ through 2 nested loops. In the internal loop, parameters $\boldsymbol{\vartheta}$ are optimised based on the log-likelihood of training data $\mathcal{D}_{train}$, resulting in a new value $\boldsymbol{\vartheta}'$ after a gradient descent update. In the external loop, the operation weights $\boldsymbol{\alpha}$ are optimised based on validation data $\mathcal{D}_{val}$ and $\boldsymbol{\vartheta}'$. This second signal can be interpreted as the reward (in Reinforcement Learning terms) or fitness (in evolutionary terms) of an architecture given optimal parameters. This algorithm repeats alternating the 2 loops until convergence:

$$\boldsymbol{\vartheta}' = \boldsymbol{\vartheta} - \xi\nabla_{\vartheta}\log p(\mathcal{D}_{train} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta}) \tag{4.4}$$

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \rho\nabla_{\alpha}\log p(\mathcal{D}_{val} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta}') \tag{4.5}$$

where ρ and ξ are scalar step sizes. By the chain rule, Equation (4.5) becomes:

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \rho\nabla_{\alpha}\log p(\mathcal{D}_{val} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta}') + \xi\nabla^2_{\alpha,\vartheta}\log p(\mathcal{D}_{train} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta})\nabla_{\vartheta'}\log p(\mathcal{D}_{val} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta}') \tag{4.6}$$

The third term contains a highly complex matrix-vector product, which is approximated by Liu et al. (2019) through the finite difference method. Since this makes training inefficient, the continuous relaxation of the architecture parameters is first estimated through Equations (4.4) and (4.5) for a smaller model with a single cell. Subsequently, given optimal $\boldsymbol{\alpha}^{\star}$, an architecture with discrete operations is derived retaining only

the operations (excluding the zero function) with the highest probability in each node. Finally, several copies of such a discretised cell are stacked into a larger model, and kept fixed while estimating the weights for each of them.

SNAS In order to avoid two stages of optimisation, as well as the cumbersome finite difference approximation, Xie et al. (2019) treat each edge operation $O^{(i,j)}$ as a categorical distribution parameterized by $\boldsymbol{\alpha}^{(i,j)}$. In order to be differentiable, each (continuously relaxed) sample is drawn through the re-parametrization trick. In particular, Xie et al. (2019) make use of the concrete distribution (Maddison et al., 2017), rewriting Equation (4.3) as:

$$\mathbf{h}^{(i)} = \sum_{o \in O} \frac{\exp((\log \alpha_o^{(i,j)} + \boldsymbol{G}_o^{(i,j)})/\lambda)}{\sum_{o' \in O} \exp((\log \alpha_{o'}^{(i,j)} + \boldsymbol{G}_{o'}^{(i,j)})/\lambda)} o(\cdot) \tag{4.7}$$

where $\boldsymbol{G}_o^{(i,j)} = -\log(-\log(\boldsymbol{U}_o^{(i,j)}))$ is a draw from the Gumbel distribution associated with $o^{(i,j)}$, and $\boldsymbol{U}_o^{(i,j)} \sim \mathcal{U}(a, b)$ is a draw from the uniform distribution. The temperature λ is steadily annealed towards 0, hence samples are one-hot vectors upon convergence.

4.3 Recasting NAS as Hierarchical Bayes

In this chapter, I propose to revisit the established NAS methods presented in Section 4.2 by recasting them as neural hierarchical Bayes. Not only this helps collocating DARTS and SNAS in the wider context of pre-neural literature, but also treats them as special cases of a more general model. This can be exploited to devise more expressive inference schemes, such as variational inference, and a different parametrisation of the dependent variables (in this case, via hyper-networks) which could improve the model performance.

As a starting point, one must note how the nested formulation of Equations (4.4) and (4.5) is highly reminiscent of gradient-based hyper-parameter optimization (Franceschi et al., 2018; Luketina et al., 2016; Maclaurin et al., 2015; Pedregosa, 2016). In fact, the architecture parameters $\boldsymbol{\alpha}$ can be interpreted as a hyper-parameter determining the cell κ. For instance, this hyper-parameter could be manually set as having any number of skip connections, layers (up to the number of nodes in the cell) and choice of activations in between. The choice of optimizing the hyper-parameters on a set of data points $\mathcal{D}_{val}$ distinct from $\mathcal{D}_{train}$ used for optimizing $\boldsymbol{\vartheta}$ is simply meant to avoid over-fitting.

Let us now concentrate on $\boldsymbol{\vartheta}'$, the weights after a gradient descent update starting from $\boldsymbol{\vartheta}$ in the 'inner loop' of the bi-level optimization of Equation (4.4). They can be

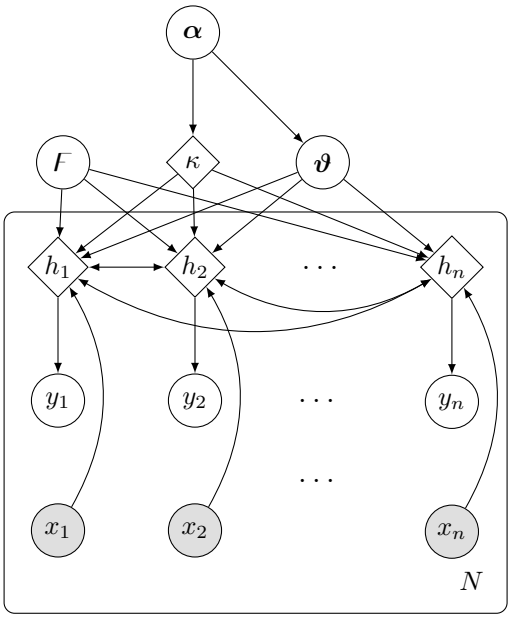


Figure 4.2 A graph of the hierarchical generative model for Neural Architecture Search.

considered as an approximation of the optimal parameters $\boldsymbol{\vartheta}^\star$ where gradient descent is truncated after a single step rather than reaching convergence:

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \rho \nabla_\alpha \log p\Big[\mathcal{D}_{val} \mid \boldsymbol{\alpha}, \underbrace{\boldsymbol{\vartheta} - \xi \nabla_\vartheta \log p(\mathcal{D}_{train} \mid \boldsymbol{\alpha}, \boldsymbol{\vartheta})}_{\approx \boldsymbol{\vartheta}^\star}\Big] \tag{4.8}$$

Then $\boldsymbol{\alpha}$ becomes a variable that constrains $\boldsymbol{\vartheta}$. Thus, the external loop of the bi-level optimization in Equation (4.5) in the original model can be recast an approximate inference over a hierarchical Bayesian model where $\boldsymbol{\vartheta}$ is integrated out (cf. Equation (2.28)):

$$\boldsymbol{\alpha}^\star = \text{argmax}_{\boldsymbol{\alpha}} \int p(\mathbf{x} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\vartheta} \tag{4.9}$$

The resulting graphical model for Neural Architecture Search is depicted in Figure 4.2. Under this formulation, both the structure of a cell κ and the appropriate parameters of the affine layers $\boldsymbol{\vartheta}$ are chosen according to a distribution $\boldsymbol{\alpha}$ over architecture parameters. For a classification task, the goal is predicting a sentence label y given some data $\mathbf{x}$ and encoder parameters F, where the encoder is any black-box function such as multilingual

BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020). This amounts to finding the joint probability $p(y, \boldsymbol{\alpha}, \boldsymbol{\vartheta} \mid \mathbf{x}, \digamma)$ as given by Equation (4.10).[1] Subsequently, the intermediate latent variables $\boldsymbol{\alpha}$ and $\boldsymbol{\vartheta}$ can be integrated out as shown in Equation (4.11) to obtain the marginal likelihood $p(y \mid \mathbf{x}, \digamma)$ of Equation (4.12).

$$p(y, \boldsymbol{\alpha}, \boldsymbol{\vartheta} \mid \mathbf{x}, \digamma) = p(y \mid \mathbf{h})\, \delta(\mathbf{h} \mid \mathbf{x}, \digamma, \boldsymbol{\alpha}, \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\alpha}) \tag{4.10}$$

$$p(y \mid \mathbf{x}, \digamma) = \int\int p(y, \boldsymbol{\alpha}, \boldsymbol{\vartheta} \mid \mathbf{x}, \digamma)\, d\boldsymbol{\vartheta}\, d\boldsymbol{\alpha} \tag{4.11}$$

$$= \int\int p(y \mid \mathbf{h})\, \delta(\mathbf{h} \mid \mathbf{x}, \digamma, \boldsymbol{\alpha}, \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\, d\boldsymbol{\vartheta}\, d\boldsymbol{\alpha} \tag{4.12}$$

Again, this formulation gives rise to intractable integrals. Therefore, we must resort to an approximate scheme such as variational inference, foreshadowed in Section 2.3.1.

This hierarchical graphical model leads to a generalisation of differentiable NAS. Firstly, the value of the architecture-dependent parameters $\boldsymbol{\vartheta}'$ does not need to be estimated necessarily through a gradient descent step as in Equation (4.4) during inference. In fact, the gradient ∇_ϑ is just a function from loss and parameters $\mathbb{R}^{1+|\boldsymbol{\vartheta}|+|\boldsymbol{\alpha}|} \to \mathbb{R}^{|\boldsymbol{\vartheta}|}$ with some special properties. This function can be substituted with any another such map (possibly with learnable parameters). Secondly, explicitly modelling priors allows for providing inducting biases and prevent catastrophic forgetting (cf. Chapter 3), in case the data shifts to a different distribution. DARTS (Liu et al., 2019) can be recovered as a special case of my general formulation by setting:

$$p(\boldsymbol{\alpha}) = \delta\{\kappa = \max(\boldsymbol{\alpha})\} \tag{4.13}$$

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{\alpha}; \xi) = \delta\{\boldsymbol{\vartheta} = \boldsymbol{\vartheta} - \xi \nabla_{\boldsymbol{\vartheta}} \log p(y \mid \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\vartheta})\} \tag{4.14}$$

where δ is a Dirac delta function.

In constructing a hierarchical generative model with the independence assumptions inherent to the graph in Figure 4.2, I assign a categorical distribution to the neural architecture variable $\kappa \sim \text{Concrete}(\boldsymbol{\alpha})$ and a Normal distribution over the variable for weight parameters responsible for affine transforms $\boldsymbol{\vartheta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}})$. Each token x_n is encoded according to the embedding parameters $p(\digamma) = \delta\{\boldsymbol{\mu}_{\digamma}\}$. Finally, it remains to establish a mechanism to sample weights conditioned on the architecture. In this experiment, I generate samples through a function $f(\cdot) : \mathbb{R}^{|\boldsymbol{\alpha}|} \to \mathbb{R}^{|\boldsymbol{\mu}_{\boldsymbol{\vartheta}}|}$, which is simply a

[1] Note that I omit κ as it results deterministically from $\boldsymbol{\alpha}$.

	PREMISE		CHOICE 1	CHOICE 2
qu	*Sipasqa cereal mikhunanpi kuruta tarirqan.*	R	*Payqa pukunman ñuqñuta churakurqan.*	*Payqa manam mikhuyta munarqanchu.*
en	The girl found a bug in her cereal.		She poured milk in the bowl.	She lost her appetite.
th	ตาของฉันแดงและบวม	C	ฉันร้องไห้	ฉันหัวเราะ
en	My eyes became red and puffy.		I was sobbing.	I was laughing.

Table 4.1 Examples of forward (Result R) and backward (Cause C) reasoning from the XCOPA validation sets.

trainable linear mapping with additional parameters that outputs weight means $\boldsymbol{\mu}_{\boldsymbol{\vartheta}}$. On the other hand, I assume the variance to be fixed: $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} = \mathbf{I}$.

The process underlying the generative model is recapped in Algorithm 1, which iterates across multiple languages $\ell \in \mathcal{T}$ and then over sentences within a language $\mathcal{D}_\ell$.

Algorithm 1 NAS Generative Model

$\kappa \sim \text{Concrete}(\boldsymbol{\alpha})$
$\boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta \leftarrow f(\boldsymbol{\alpha})$
$\boldsymbol{\vartheta} \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta)$
$F \sim \delta\{\boldsymbol{\mu}_F\}$
for $\ell \in \mathcal{T}$
  for $s \in \mathcal{D}_\ell$
    $\mathbf{x}^{(s)} \leftarrow \text{APPLY}(F, x_1^{(s)}, \ldots, x_n^{(s)})$
    $\mathbf{h}^{(s)} \leftarrow \text{APPLY}(\kappa, \boldsymbol{\vartheta}, \mathbf{x}^{(s)})$
    $y^{(s)} \sim p(\cdot \mid \mathbf{h}^{(s)})$

4.4 Multilingual Commonsense Reasoning

After the definition of Neural Architecture Search via hierarchical Bayes in Section 4.3, in what follows I devise an experiment to evaluate how it compares to state-of-the-art methods in terms of both performance and sample efficiency. Note, however, that it is already possible to outline an advantage of the proposed method, its run-time efficiency, on theoretical grounds. In fact, rather than grid searching different hyper-parameter configurations, optimising architecture parameters requires just a single run. In this section, I present the dataset I created for the experiment, whereas the following

Section 4.5 elaborates on the experimental setup, providing details on the neural model and the inference scheme.

The ideal benchmark to evaluate whether Hierarchical Bayes NAS favours sample-efficient natural language understanding i) should be a genuinely diverse multilingual dataset, where the internal variety of typological features is privileged over the abundance of digital resources in each language; ii) should require the transfer of high-level, abstract knowledge in order to be solved successfully. Admittedly, there already exist a few natural language understanding datasets that satisfy (i), such as TyDiQA (Clark et al., 2020) for passage-based question answering or XNLI (Conneau et al., 2018) for natural language inference. However, all of these mostly rely on explicit textual information, and are therefore not really suitable for testing the transfer of implicit knowledge.

A perfect candidate for the requirement (ii) instead is *commonsense reasoning*, a critical component of any natural language understanding system (Davis and Marcus, 2015). In fact, commonsense reasoning must bridge between premises and possible hypotheses with *world knowledge* that is not explicit in text (Singer et al., 1992). Such world knowledge encompasses, among other aspects: temporal and spatial relations, causality, laws of nature, social conventions, politeness, emotional responses, and multiple modalities. Hence, it corresponds to the individuals' expectations about typical situations (Shoham, 1990). Moreover, there are often multiple legitimate chains of sentences that can be invoked in between premises and hypotheses. In short, commonsense reasoning does not just involve understanding what is possible, but also ranking what is most *plausible*.

A seminal work on the quantitative evaluation of commonsense reasoning is the Choice Of Plausible Alternatives dataset (COPA; Roemmele et al., 2011), which focuses on cause–effect relationships. In recent years, more datasets have been dedicated to other facets of world knowledge (Bhagavatula et al., 2020; Bisk et al., 2020b; Rashkin et al., 2018; Sakaguchi et al., 2020; Sap et al., 2019, *inter alia*). Unfortunately, the extensive efforts related to this thread of research have so far been limited only to the English language.[2] Such a narrow scope not only curbs the development of natural language understanding tools in other languages (Bender, 2011; Ponti et al., 2019a), but also exacerbates the Anglo-centric bias in modelling commonsense reasoning. In fact, the expectations about typical situations do vary across cultures (Thomas, 1983).

In order to fill the gap of a multilingual dataset for commonsense reasoning, I develop a novel dataset, XCOPA (see examples in Table 4.1), by carefully translating and re-

[2]The only exception is direct translation of the 272 paired English Winograd Schema Challenge instances to Japanese (Shibata et al., 2015), French (Amsili and Seminck, 2017), and Portuguese (Melo et al., 2020).

	Range	XCOPA	TyDiQA	XNLI	XQUAD	MLQA	PAWS-X
Typology	[0, 1]	0.41	0.41	0.39	0.36	0.32	0.31
Family	[0, 1]	1	0.9	0.5	0.6	0.66	0.66
Geography	[0, ln 6]	1.67	0.92	0.37	0	0	0

Table 4.2 Indices of typological, genealogical, and areal diversity for the language samples of a set of NLU datasets.

annotating the validation and test sets of English COPA into 11 target languages from 11 distinct families, and 4 geographical macro-areas. The key design goals are: i) to align examples across languages in order to make performance scores comparable; ii) to ensure high quality, naturalness and idiomaticity of each monolingual dataset. In the following sections, I outline the criteria underlying the selection of languages and the guidelines adopted to achieve the above-mentioned goals.

4.4.1 Language Sampling

Multilingual evaluation benchmarks assess the *expected performance* of a model across languages. However, should such languages be sampled according to the distribution of digital texts or rather based on the distribution over the languages spoken around the world? The former strategy is unreliable, as languages rich in resources tend to belong to the same families and areas, which facilitates knowledge transfer and hence leads to an overestimation of the expected performance (Gerz et al., 2018b; Ponti et al., 2019a).

Moreover, rather than samples that account for independent and identically distributed draws from the 'true' language distribution (known as *probability* sampling), I opt for a *uniform* distribution of linguistic phenomena, which encourages the inclusion of outliers (known as *variety* sampling; Dryer, 1989; Rijkhoff et al., 1993). Thus, the performance on XCOPA also reflects the *robustness* of a model, i.e. its resilience to phenomena that are unlikely to be observed in the training data.

Inspired by Rijkhoff et al. (1993) and Miestamo (2004), I propose a series of simple and interpretable metrics that quantify diversity of a language sample independent of its size: **1)** a *typology* index based on 103 typological features of each language from URIEL (Littell et al., 2017), originally sourced from the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013). Each feature is binary and indicates the presence or absence of a phenomenon in a language. I estimate the entropy of the distribution of values in a sample. Afterwards, I average across all 103 feature-specific entropies.

	ET	HT	ID	IT	QU	SW	TA	TH	TR	VI	ZH
val	97.0	97.0	99.0	98.0	98.0	99.0	100.0	99.0	97.0	97.0	96.0
test	98.2	96.4	100.0	97.0	94.8	99.0	98.6	98.2	96.4	98.4	96.6

Table 4.3 Percentage of annotated labels in each language agreeing with the majority label. Note that the majority label is highly reliable, as I observed a 100% agreement with the development set labels in the original COPA.

Intuitively, if all values are equally represented, the entropy is high. If all languages have identical features, the entropy is 0; **2)** The *family* index is simply the number of distinct families divided by the sample size; **3)** The *geography* index is the entropy of the distribution over macro-areas in a sample.[3]

The sample of languages for XCOPA aims at maximising these indices. In particular, XCOPA includes Estonian (ET), Haitian Creole (HT), Indonesian (ID), Italian (IT), Cusco-Collao Quechua (QU),[4] Kiswahili (SW), Tamil (TA), Thai (TH), Turkish (TR), Vietnamese (VI), and Mandarin Chinese (ZH). These languages belong to distinct families, respectively: Uralic, Creole, Austronesian, Indo-European, Yuman–Cochimí, Niger-Congo, Dravidian, Kra-Dai, Turkic, Austroasiatic, and Sino-Tibetan. Moreover, HT and QU are spoken in Central and South America, respectively, which are underrepresented macro-areas. I report the 3 metrics in Table 4.2 and compare them to samples from other standard multilingual NLU datasets. XCOPA offers the most diverse sample in terms of typology (on a par with TyDiQA), family, and geography.

4.4.2 Annotation Procedure

As shown in Table 4.1, each XCOPA instance corresponds to a premise, a question ("*What was the* CAUSE?" or "*What happened as a* RESULT?"), and two alternatives. The task is framed as binary classification where the machine has to predict the more plausible choice. For each target language, XCOPA comprises 100 annotated data instances in the validation set and 500 instances as the test set, which are translations from the respective English COPA validation and test set, see Table 4.1 again. Translators performed labelling prior to translation, deciding on the correct alternative for the English premise and preserving the correctness of the same alternative in translation. I measure

[3]Six macro-areas, as defined by Dryer (1989), are Africa, Eurasia, Southeast Asia and Oceania, Australia and New Guinea, North America, and South America.

[4]The translator is an Eastern Apurímac Quechua speaker.

inter-translator agreement using the Fleiss' κ statistic (Fleiss, 1971): the obtained scores of 0.921 for development data and 0.911 for test data reveal very high agreement between translators. In fact, Landis and Koch (1977) define $\kappa \geq 0.81$ as almost perfect agreement.

From the 11 sets of annotation labels, I obtain the majority labels (i.e., 6+ translators agree). I observe perfect agreement between these majority labels and the English COPA labels for development data. I then compute the percentage of annotated labels which agree with the majority label for each language individually, reported in Table 4.3, and find very high agreement across 11 languages. On average, 2.1% of labels in the validation set and 2.4% of labels in the test set do not match the majority label.

The choice of translating from English, rather than creating novel instances, abides by the principle of maintaining examples aligned across languages. While the commonly used translation approach achieves this objective, however, it is prone to compromise the idiomaticity, bending the target language to the structural and lexical properties of the source language To avoid these pitfalls, I adopt guidelines that address language-specific challenges, such as the lack of equivalent concepts or the grammatical expression of tense and aspect.

In particular, the scenarios included in English COPA were authored by American English speakers with a particular cultural background. It is therefore inevitable that some concepts, intended as commonplace, sound unusual or even completely foreign in the target language. Examples include: (i) concrete referents with no language-specific term available (e.g., *bowling ball*, *hamburger*, *lottery*); (ii) systems of social norms absent in the target culture, e.g., traffic regulations (e.g., *parallel parking*, *parking meter*); (iii) social, political, and cultural institutions and related terminology (e.g., e.g., *mortgage*, *lobbyist*, *gavel*); (iv) idiomatic expressions (e.g., *put the caller on hold*).

In such cases, the translators were advised to resort to (i) paraphrasing; (ii) substitutions with similar concepts, e.g., 'faucet' is replaced with 'pipe' in Tamil (குழாய், *kuḻāy*) and Haitian Creole (*tiyo*); or (iii) phonetically transcribed loan words, e.g., in Tamil: பௌலிங் பந்து (*pauliṅ pantu*, 'bowling ball'), சோப்பு (*cōppu*, 'soap').

An in-depth analysis revealed the source of inter-translator disagreement on the validation set annotations across languages.[5] In only two cases of discrepancy did the translator's cultural frame of reference play a role. For instance, one example required to be acquainted with American court trials: *The judge pounded the gavel.* CAUSE: (a) *The courtroom broke into uproar.* (b) *The jury announced its verdict.* Most disagreement cases (87.5%), however, seem to be culturally independent and concern genuinely ambiguous

[5]Overall, there were 10 validation set questions with 1 translator out of 11 in disagreement, 5 questions with 2, and 1 question with 3.

cases (e.g. *The detective revealed an anomaly in the case.* RESULT: (a) *He finalized his theory.* (b) *He scrapped his theory.*).

4.5 Experimental Setup

I will now outline the setup of the experiment to perform multiple-choice classification on XCOPA for both a baseline based on pre-training and fine-tuning and the proposed model based on Hierarchical Bayes Neural Architecture Search (HBNAS).

Multiple-Choice Classification. XCOPA is a multiple-choice classification task: given a premise and a prompt (CAUSE or RESULT), the goal is to select the more plausible of the two answer choices (see Table 4.1). Due to training data scarcity in COPA, I probe the usefulness of first "pretraining" the classifier on larger multiple-choice English commonsense reasoning datasets, and specifically SOCIALIQA (SIQA; Sap et al., 2019). As different multiple-choice selection tasks differ in the number of choices (e.g., there are 2 possible answers in XCOPA, whereas there are 3 in SIQA), a classifier with a fixed number of classes is not a good fit for this scenario. I thus follow Sap et al. (2019) and couple the (pretrained) encoder with a feed-forward network which produces a single scalar score for each of the possible answers. The scores for individual choices are then concatenated and passed to the softmax function. Besides the standard state-of-the-art transfer models based on pretraining and fine-tuning, I also benchmark the HBNAS model and measure how it fares against these competitive baselines.

Encoder Model. I evaluate the following state-of-the-art pretrained multilingual encoders: **1)** multilingual BERT (MBERT) (Devlin et al., 2019) and XLM-on-RoBERTa (Conneau et al., 2020), both the Base (XLM-R) and Large (XLM-R-L) variant, in the standard *fine-tuning regime* (i.e., their parameters are fine-tuned together with the task classifier's parameters), and **2)** multilingual Universal Sentence Encoder (USE) (Yang et al., 2019) in the *feature-based regime* (i.e., its parameters are fixed during the task classifier's training). Both MBERT and XLM-R include all XCOPA languages in their pretraining data spanning ∼100 languages, except for Haitian Creole and Quechua. Multilingual USE was trained on 16 languages, covering IT, TH, TR, and ZH from the XCOPA language sample. The hidden state size H of each encoder is equivalent to the configuration of the pre-trained model: multilingual BERT (Base, $H = 768$), XLM-R (Base, $H = 768$; Large, $H = 1,024$), and multilingual USE (Large, $H = 512$).

	Train dataset		Model selection	
Setup	SIQA	COPA	EN	target
CO-ZS		✓	✓	
CO-TLV		✓		✓
SI-ZS	✓		✓	
SI+CO-ZS	✓	✓	✓	
SI+CO-TLV	✓	✓		✓

Table 4.4 Different fine-tuning and transfer setups. CO=COPA; SI=SIQA; ZS=Zero-Shot; TLV=Target Language Validation (Set).

Encoder Input. For each instance, I couple each of the answer choices with the concatenation of the premise and the prompt and feed that as a "sentence pair" input to MBERT and XLM-R, or as a single "sentence" to USE.[6]

Classifier Head. Let c_i be the i-th answer choice of an instance of multiple-choice dataset (i.e., $i \in \{1, 2\}$ in COPA and $i \in \{1, 2, 3\}$ in SIQA) and let $\mathbf{x}_i \in \mathbb{R}^H$ (with H as the vector size of the encoder) be the encoding of its corresponding input consisting of the premise, prompt and the answer itself, as explained above.[7] The predicted score $\hat{y}_i$ for the answer c_i is then obtained with a L-layer feed-forward network. I obtain the score $\hat{y}_i$ for each answer c_i and concatenate them into a prediction vector to which I apply softmax normalisation: $\hat{\mathbf{y}} = \text{softmax}([\hat{y}_1, \ldots, \hat{y}_N])$, where N is the number of answers in the multiple-choice selection dataset. The loss for the training instance is then the standard cross-entropy classification loss.

Transfer Learning Setups. I evaluate each model in different transfer learning setups based on **1)** different sources of training data: SIQA,[8] COPA, or both; and **2)** different model selection regimes for hyper-parameter tuning and early stopping (based on English or target language validation set). The resulting combinations are shown in Table 4.4.

Hyper-parameters. The maximum sequence length of input sentences is fixed to 64 tokens. Training runs with an effective batch size of 32, for 5 epochs. Weight parameters

[6]For MBERT and XLM-R, I insert the standard special tokens. For example, for the last example from Table 4.1 and Choice 1, the input for MBERT would be as follows: '*[CLS] My eyes became red and puffy. What was the cause? [SEP] I was sobbing. [SEP]*'.

[7]For MBERT and XLM-R $\mathbf{x}_i$ is the Transformer representation of the sequence start token. For USE, $\mathbf{x}_i$ is the average of contextualised vectors of all tokens.

[8]The SIQA dataset is similar in nature to COPA (i.e., it is a multiple-choice dataset for commonsense reasoning about social interactions, with open-format prompts and three answer choices). It comes with a much larger training set, consisting of 33K instances and therefore can provide useful learning signal also for causal commonsense reasoning in XCOPA.

$\boldsymbol{\vartheta}$ optimised through an Adam optimiser (Kingma and Ba, 2015) a learning rate of 8×10^{-6}. Gradients are clipped to a norm of 1.

When performing HBNAS, architecture parameters $\boldsymbol{\alpha}$ are updated through a separate Adam optimiser with a learning rate of 3×10^{-3} and a weight decay of 10^{-3}. Temperature τ is annealed from 1 to 0 with a linear schedule. During variational inference, I assume a prior of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for the neural weights $\boldsymbol{\vartheta}$ and a Dirichlet distribution $\mathrm{Dir}(\mathbf{1})$ for the architecture $\boldsymbol{\alpha}$. The classifier depth with the highest validation performance was 2 for the baseline, and 4 for NAS. The baseline classifier is a Feed-forward network with a *tanh* activation and no skip connections.

Zero-Shot and Few-Shot Learning. In the zero-shot learning setting, the model trained on English is directly evaluated on the multi-lingual XCOPA benchmark. Instead, in the few-shot learning setting, the model is further trained on XCOPA development set individually for each target language. Since no further development data is available to grid search the hyper-parameters, those found on the COPA English developments set are retained.

4.6 Results

In this section, I provide the results for different transfer learning settings (training and validation data) and choices of encoders. Subsequently, I adopt the best model as a baseline for HBNAS and compare their performances.

4.6.1 Choice of Encoder and Transfer Setting

Table 4.5 shows the aggregate accuracy of MBERT, XLM-R and USE over 11 XCOPA languages for each of the previously described training setups from Table 4.4. Comparing the XCOPA results with the English COPA performance of the monolingual English BERT (Base) reported by Sap et al. (2019), it immediately emerges that even the best setting in XCOPA yields a drop of -7% (from an accuracy of ~63) with COPA-only fine-tuning and -17% (from an accuracy of ~80) with SIQA and COPA fine-tuning. This reinforces recent suspicions (Cao et al., 2020; Hu et al., 2020) that massively multilingual pretrained transformers do not offer a completely satisfactory solution for language transfer.

XLM-R outperforms MBERT and USE in all setups, but the gains are pronounced only in setups in which the models were first fine-tuned on SIQA (SI-ZS, SI+CO-ZS, and SI+CO-TLV). USE outperforms MBERT surprisingly often. This might have been

Setup	Model	All	MBERT ∩ XCOPA	USE ∩ XCOPA
CO-ZS	XLM-R	55.6	56.9	55.4
	XLM-R-L	52.4	52.5	52.1
	MBERT	54.1	54.4	55.7
	USE	54.7	56.0	58.1
CO-TLV	XLM-R	55.1	56.4	55.2
	XLM-R-L	51.6	51.7	52.1
	MBERT	54.2	54.5	55.8
	USE	54.8	55.4	59.0
SI-ZS	XLM-R	60.1	62.3	62.9
	XLM-R-L	68.4	72.1	72.9
	MBERT	54.7	55.6	56.4
	USE	55.0	56.4	60.1
SI+CO-ZS	XLM-R	59.0	60.7	61.9
	XLM-R-L	67.3	70.8	71.8
	MBERT	55.8	56.8	57.9
	USE	54.1	54.9	58.9
SI+CO-TLV	XLM-R	60.7	63.5	63.6
	XLM-R-L	**69.1**	**72.8**	**74.6**
	MBERT	54.4	54.8	54.2
	USE	54.3	55.2	59.1

Table 4.5 Summary of XCOPA results. **All**: average over all 11 XCOPA languages; **MBERT ∩ XCOPA**: average over 9 XCOPA languages (without HT and QU) included in MBERT and XLM-R pretraining; **USE ∩ XCOPA**: average over 4 XCOPA languages (IT, TH, TR, and ZH), included in the USE pretraining.

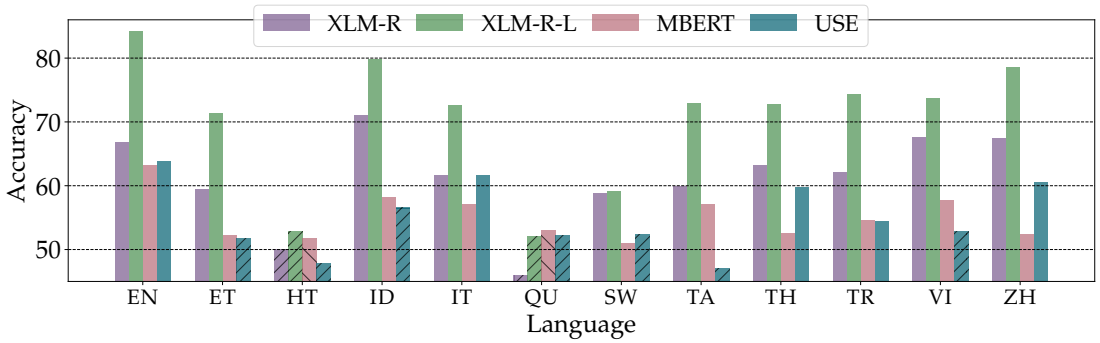


Figure 4.3 Per-language XCOPA results for XLM-R, MBERT, and USE in the SIQA + COPA-TLV setup. Striped bars correspond to language-model pairs where the language was not included in model pretraining.

	Model	EN	ET	ID	IT	SW	TA	TH	TR	VI	ZH	all
Zero	XLM-R Large 2	85.4	**71.8**	80.6	75.6	62.0	70.8	72.2	73.6	76.8	78.2	74.70
	XLM-R Large 4	83.8	**71.8**	79.6	73.2	64.2	**73.8**	71.0	71.2	75.2	**82.2**	74.60
	HBNAS 4	**86.0**	70.2	**82.6**	**75.8**	**65.2**	72.4	**74.8**	**75.4**	**78.0**	81.0	**76.14**
Few	XLM-R Large 2	**86.8**	70.8	**84.8**	78.0	60.8	72.0	74.8	73.0	78.6	**82.2**	76.18
	XLM-R Large 4	86.4	72.6	80.6	74.2	64.4	71.2	74.2	74.2	77.4	81.2	75.64
	HBNAS 4	86.4	**74.4**	81.6	**79.6**	**64.8**	**75.2**	**76.6**	**76.0**	**79.0**	81.8	**77.54**

Table 4.6 Zero-shot and few-shot results on XCOPA comparing NAS to the best baseline. Numbers after a model name indicate the depth of the classifier on top of the encoder.

expected in the COPA-only setups (CO-ZS and CO-TLV) where the small COPA training set is insufficient to meaningfully fine-tune MBERT transformer parameters. However, the finding that MBERT does not benefit more than USE from prior SIQA training is surprising and warrants further investigation. What is more, USE in some setups even outperforms MBERT for some of the languages (e.g., ID, TA, SW) on which MBERT was pretrained and USE was not (cf. the scores in the MBERT $\cap$ XCOPA column). I speculate that this is due to the combination of two effects: (1) the infamous "curse of multilinguality" (Conneau et al., 2020) is much more pronounced for MBERT (which is pretrained on 104 languages) than for USE, pretrained on only 16 languages; and (2) the presence of subword-level similarities between XCOPA target languages and the 16 languages used in USE pretraining. Unsurprisingly, the Large XLM-R substantially outperforms its Base counterpart in all setups with SIQA training. Because of almost 3 times more parameters (355M vs. 125M), XLM-R-L stores more language-specific information for each pretraining language. The large parameter space, however, also causes XLM-R-L to underperform XLM-R in COPA-only setups (CO-ZS and CO-TLV), when exposed only to a tiny COPA fine-tuning dataset.

Also note that training models only on SIQA yields performance that is comparable (and for MBERT and USE often better) to the performance I obtain with additional COPA training (setups SI + CO-ZS and SI + CO-TLV). While this is in part due to the limited size of the COPA training set, it confirms the assumption that SIQA and COPA are highly compatible tasks. Moreover, only slight gains are achieved by hyper-parameter tuning on the target language validation set (TLV).

Figure 4.3 shows per-language performance in the best setup, SIQA + COPA-TLV, while I provide detailed results for all other setups in Table C.1 in appendix. As expected,

all models fluctuate around random-level performance on out-of-sample languages, HT and QU. For all other languages, XLM-R outperforms MBERT. Surprisingly, I also observe that for some languages (ID, VI, ZH) performance of transfer from English is slightly higher than the actual performance in English, without transfer. Another observation is that the transfer performance is often better for some languages typologically distant from English than for languages closer to English (e.g., TH, VI, ZH versus IT). This might be partially due to superior representations of languages such as ZH and TH in the pretrained models due to their large training data and very specific scripts (input embedding parameters do not need to be shared with other languages).

4.6.2 Effectiveness of HBNAS

Given the results of Section 4.6.1, I adopt the best transfer setting and encoder as a baseline for the proposed model, HBNAS. In particular, the baseline consists of an XML-R encoder and a 2-layer MLP classifier, both trained on SIQA. Early stopping is based on the XCOPA validation sets of all target languages. Such baseline is compared to an equivalent model that learns the classifier architecture through NAS as detailed in Section 4.3. To make the comparison as fair as possible, I also report the scores for a baseline with an identical number of parameters, i.e. with 4 classifier layers. Finally, I omit the results of DARTS for brevity: I verified that while it achieves performance gains that are not statistically significant compared to the other baselines, it incurs excessively more training time.

Results are shown in Table 4.6. Neural architecture search achieves superior results for both zero-shot and few-shot learning. In particular, it outperforms the baseline in 7/10 languages (except ET, TA, and ZH) in the zero-shot setting and in 7/10 languages (except EN, ID, and ZH) in the few-shot setting. Note that the gains are especially evident in languages whose scores lie on the low side of the spectrum, thus making the method especially suited in resource-lean scenarios. In average, HBNAS obtains improvements of 1.44 points in accuracy for zero-shot learning, and 1.24 points in accuracy for few-shot learning. These figures confirm the hypothesis that the proposed NAS method can help construct a prior over both parameters and architectures that enables the transfer of high-level, abstract knowledge necessary for causal common-sense reasoning.

Such prior is visualised in Figure 4.5, where I plot the heat map of the learned posterior for $\boldsymbol{\alpha}$. These values translate into the cell structure illustrated in Figure 4.1.

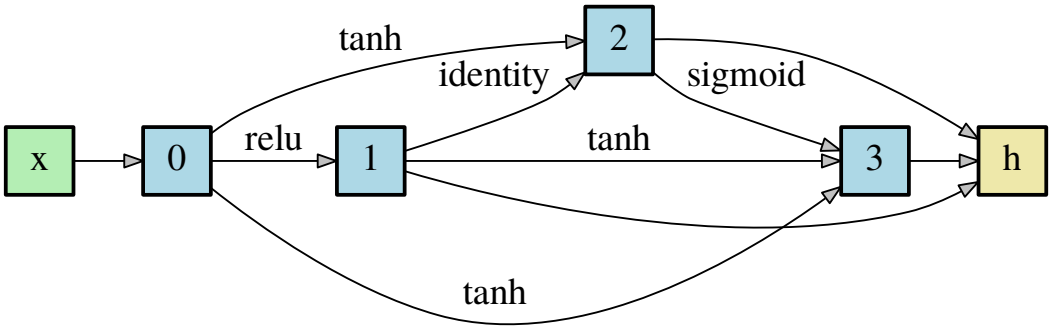


Figure 4.4 Learned cell structure.

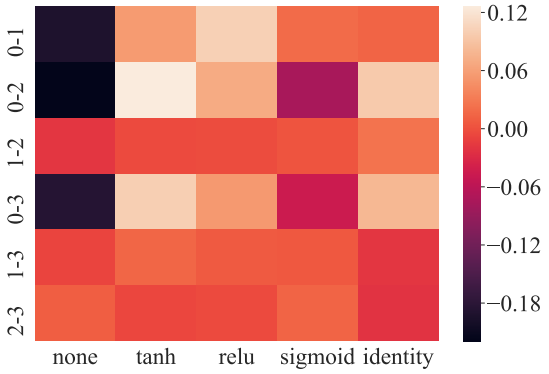


Figure 4.5 Heatmap of the $\boldsymbol{\alpha}$ logits.

4.7 Conclusions

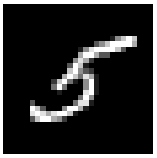
In this chapter, I have explored the idea of constructing a prior over both neural weights and architectures (encompassing layer connections and nonlinear activations) in order to facilitate zero-shot and few-shot natural language understanding in novel languages. To do so, I have recast neural architecture search as implicitly defining a hierarchical Bayesian model. Moreover, I have shown how to adapt this framework to state-of-art algorithms for natural language processing, and in particular Transformer-based encoders pretrained on language modelling.

As a challenging benchmark to evaluate the proposed method, I have created the Cross-lingual Choice of Plausible Alternatives (XCOPA) dataset for causal commonsense

reasoning. All XCOPA instances are aligned across 11 languages, which enables cross-lingual comparisons. The language selection was informed by variety sampling, in order to maximise diversity in terms of typological features, geographical macro-area, and language family. This allows for testing the robustness of machine learning models for an array of rare phenomena displayed by the chosen languages.

To establish strong baselines on this dataset, I ran a series of cross-lingual transfer experiments, evaluating state-of-the-art transfer methods based on multilingual pretraining and fine-tuning on English. I observed that, although these methods perform better than chance, they still lag significantly behind the monolingual supervised learning setting. Overall, the scores are held down by the 'curse of multilinguality', the need to account for a wide sample of languages in pretraining. In addition, the transfer seems not to depend that much on the distance from the source, but rather on the abundance of target language data in multilingual pretraining.

Leveraging neural architecture search on the classifier network yielded additional gains in both the zero-shot and few-shot learning settings, thus demonstrating the ability of the learned prior to capture causal world knowledge. These results hold promise to foster further research in multilingual commonsense reasoning and cross-lingual transfer, and possibly apply the novel method to other multi-lingual tasks such as question answering (Clark et al., 2020) or natural language inference (Conneau et al., 2018).

Chapter 

Modular Design via Parameter Factorisation

5.1 Introduction

The annotation efforts in NLP have achieved impressive feats, such as the Universal Dependencies (UD) project (Nivre et al., 2019) which now includes 83 languages. But, even UD covers only a meagre subset of the world's estimated 8,506 languages (Hammarström et al., 2020) (cf. Section 2.2.3). Moreover, the Association for Computational Linguistics Wiki[1] lists 24 separate NLP tasks. Labelled data, which is both costly and labour-intensive, is missing for many of such task–language combinations. This shortage hinders the development of computational models for the majority of the world's languages (Ponti et al., 2019a; Snyder and Barzilay, 2010).

As argued in Section 2.2.4, a common solution is transferring knowledge across domains, such as tasks and languages (Talmor and Berant, 2019; Yogatama et al., 2019), which holds promise to mitigate the lack of training data inherent to a large spectrum of NLP applications (Agić et al., 2016; Ammar et al., 2016; Ponti et al., 2018a; Täckström et al., 2012; Ziser and Reichart, 2018, *inter alia*). In the most extreme scenario, *zero-shot learning*, no annotated examples are available for the target domain. In particular, zero-shot transfer across *languages* implies a change in the data domain, and leverages information from resource-rich languages to tackle the same task in a previously unseen target language (Artetxe and Schwenk, 2019; Lin et al., 2019; Ponti et al., 2019a; Rijhwani et al., 2019, *inter alia*). Zero-shot transfer across *tasks* within the same language (Ruder et al., 2019a), on the other hand, implies a change in the space of labels.

[1] aclweb.org/aclwiki/State_of_the_art

As the main contribution of this chapter, I propose a Bayesian generative model of the neural parameter space (Ponti et al., 2021). I assume that this space is structured, and for this reason factorisable into task- and language-specific latent variables.[2] By performing transfer of knowledge from both related tasks *and* related languages (i.e., from *seen* combinations), my model allows for zero-shot prediction on *unseen* task–language combinations. For instance, the availability of annotated data for part-of-speech (POS) tagging in Wolof and for named-entity recognition (NER) in Vietnamese supplies plenty of information to infer a task-agnostic representation for Wolof and a language-agnostic representation for NER. Conditioning on these, the appropriate neural parameters for Wolof NER can be generated at evaluation time. While this idea superficially resembles matrix completion for collaborative filtering (Dziugaite and Roy, 2015; Mnih and Salakhutdinov, 2008), the neural parameters are latent and are non-identifiable. Rather than recovering missing entries from partial observations, in my approach I reserve separate latent variables to each language and each task to tie together neural parameters for combinations that have either of them in common.

I adopt a Bayesian perspective towards inference. The posterior distribution over the model's latent variables is approximated through stochastic variational inference (SVI; Hoffman et al., 2013). Given the enormous number of parameters, I also explore a memory-efficient inference scheme based on a diagonal plus low-rank approximation of the covariance matrix. This guarantees that the model remains both expressive and tractable.

I evaluate the model on two sequence labelling tasks: POS tagging and NER, relying on a typologically representative sample of 33 languages from 4 continents and 11 families. The results clearly indicate that the generative model surpasses standard baselines based on cross-lingual transfer 1) from the (typologically) nearest source language; 2) from the source language with the most abundant in-domain data (English); and 3) from multiple source languages, in the form of either a multi-task, multi-lingual model with parameter sharing (Wu and Dredze, 2019) or an ensemble of task- and language-specific models (Rahimi et al., 2019).

Finally, I empirically demonstrate the importance of modelling uncertainty during inference through Monte Carlo approximations of Bayesian model averaging, as opposed to point estimates. While yielding comparable performances, this endows neural networks with the ability to "fail loudly" (Rabanser et al., 2019) in low-confidence settings such as zero-shot cross-lingual and cross-task transfer. In particular, whenever the posterior

[2]By latent variable I mean every variable that has to be inferred from observed (directly measurable) variables. To avoid confusion, I use the terms *seen* and *unseen* when referring to different task–language combinations.

predictive distributions in a domain display a high entropy, prediction is such domain can be avoided. As a result, the generative model enhances both accuracy and robustness in low-resource NLP tasks.

5.2 Bayesian Generative Model

In this chapter, I propose a Bayesian generative model for multi-task, multi-lingual NLP. I train a single Bayesian neural network for several tasks and languages jointly. Formally, I consider a set $T = \{t_1, \ldots, t_n\}$ of n tasks and a set $L = \{l_1, \ldots, l_m\}$ of m languages. The core modelling assumption I make is that the parameter space of the neural network is *structured*: specifically, I posit that certain parameters correspond to tasks and others correspond to languages. This structure assumption allows us to generalise to unseen task–language pairs. In this regard, the model is reminiscent of matrix factorisation as applied to collaborative filtering (Dziugaite and Roy, 2015; Mnih and Salakhutdinov, 2008).

I now describe the generative model in three steps that match the nesting level of the plates in the diagram in Figure 5.1. Equivalently, the reader can follow the nesting level of the **for** loops in Algorithm 2 for an algorithmic illustration of the generative story.

(1) **Sampling Task and Language Representations:** To kick off the generative process, I first sample a latent representation for each of the tasks and languages from multivariate Gaussians: $\boldsymbol{t}_i \sim \mathcal{N}(\boldsymbol{\mu}_{t_i}, \boldsymbol{\Sigma}_{t_i}) \in \mathbb{R}^h$ and $\boldsymbol{l}_j \sim \mathcal{N}(\boldsymbol{\mu}_{l_j}, \boldsymbol{\Sigma}_{l_j}) \in \mathbb{R}^h$, respectively. While I present the model in its most general form, I take $\boldsymbol{\mu}_{t_i} = \boldsymbol{\mu}_{l_j} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{t_i} = \boldsymbol{\Sigma}_{l_j} = \boldsymbol{I}$ for the experimental portion of this chapter.

(2) **Sampling Task–Language-specific Parameters:** Afterwards, to generate task–language-specific neural parameters, we sample $\boldsymbol{\vartheta}_{ij}$ from $\mathcal{N}(f_{\boldsymbol{\psi}}(\boldsymbol{t}_i, \boldsymbol{l}_j), \text{diag}(f_{\boldsymbol{\phi}}(\boldsymbol{t}_i, \boldsymbol{l}_j))) \in \mathbb{R}^d$ where $f_{\boldsymbol{\psi}}(\boldsymbol{t}_i, \boldsymbol{l}_j)$ and $f_{\boldsymbol{\phi}}(\boldsymbol{t}_i, \boldsymbol{l}_j)$ are learned deep feed-forward neural networks $f_{\boldsymbol{\psi}} : \mathbb{R}^h \to \mathbb{R}^d$ and $f_{\boldsymbol{\phi}} : \mathbb{R}^h \to \mathbb{R}^d_{\geq 0}$ parametrized by $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$, respectively, similar to Kingma and Welling (2014). These transform the latent representations into the mean $\boldsymbol{\mu}_{\theta_{ij}}$ and diagonal of the covariance matrix $\boldsymbol{\sigma}^2_{\theta_{ij}}$ for the parameters $\boldsymbol{\vartheta}_{ij}$ associated with t_i and l_j. The feed-forward network $f_{\boldsymbol{\psi}}$ just has a final linear layer as the mean can range over $\mathbb{R}^d$ whereas $f_{\boldsymbol{\phi}}$ has a final softplus (defined in Section 5.3) layer to ensure it ranges only over $\mathbb{R}^d_{\geq 0}$. Following Stolee and Patterson (2019), the networks $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ take as input a linear function of the task and language vectors: $\boldsymbol{t} \oplus \boldsymbol{l} \oplus (\boldsymbol{t} - \boldsymbol{l}) \oplus (\boldsymbol{t} \odot \boldsymbol{l})$, where $\oplus$ stands for concatenation and $\odot$ for element-wise multiplication. The sampled neural parameters $\boldsymbol{\vartheta}_{ij}$ are partitioned into a weight

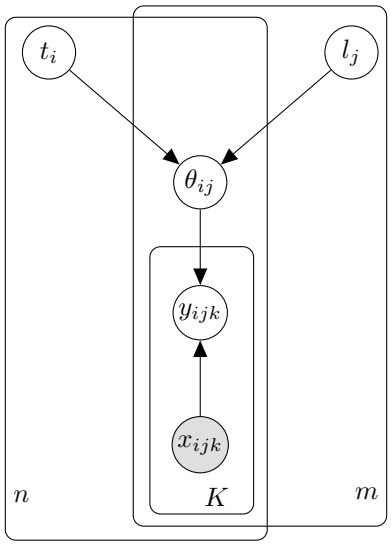


Figure 5.1 Graph (plate notation) of the generative model based on parameter space factorisation. Shaded circles refer to observed variables.

$\boldsymbol{W}_{ij} \in \mathbb{R}^{e \times c}$ and a bias $\boldsymbol{b}_{ij} \in \mathbb{R}^{c}$, and reshaped appropriately. Hence, the dimensionality of the Gaussian is chosen to reflect the number of parameters in the affine layer, $d = e \cdot c + c$, where e is the dimensionality of the input token embeddings (detailed in the next paragraph) and c is the maximum number of classes across tasks.[3] The number of hidden layers and the hidden size of $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ are hyper-parameters discussed in Section 5.4.2. I tie the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ for all layers except for the last to reduce the parameter count. I note that the space of parameters for all tasks and languages forms a tensor $\Theta \in \mathbb{R}^{n \times m \times d}$, where d is the number of parameters of the largest model.

(3) **Sampling Task Labels:** Finally, the k^{th} label y_{ijk} for the i^{th} task and the j^{th} language is sampled from a final softmax: $p(y_{ijk} \mid \mathbf{x}_{ijk}, \boldsymbol{\vartheta}_{ij}) = \text{softmax}(W_{ij}\, \textsc{bert}(\mathbf{x}_{ijk}) + \boldsymbol{b}_{ij})$ where $\textsc{bert}(\mathbf{x}_{ijk}) \in \mathbb{R}^{e}$ is the multi-lingual BERT (Pires et al., 2019) encoder. The incorporation of m-BERT as a pre-trained multilingual embedding allows for enhanced cross-lingual transfer.

Consider the Cartesian product of all tasks and languages $T \times L$. We can decompose this product into seen task–language pairs $\mathcal{S}$ and unseen task–language pairs $\mathcal{U}$, i.e. $T \times L = \mathcal{S} \sqcup \mathcal{U}$. Naturally, we are only able to train the model on the seen task–language

[3] Different tasks might involve different class numbers, the number of parameters hence oscillates. The extra dimensions not needed for a task can be considered as padded with zeros.

Algorithm 2 Generative Model of Neural Parameters for Multi-task, Multi-lingual NLP.

for $t_i \in T$
  $\boldsymbol{t}_i \sim \mathcal{N}(\boldsymbol{\mu}_{t_i}, \boldsymbol{\Sigma}_{t_i})$
for $l_j \in L$
  $\boldsymbol{l}_j \sim \mathcal{N}(\boldsymbol{\mu}_{l_j}, \boldsymbol{\Sigma}_{l_j})$
for $t_i \in T$
  for $l_j \in L$
    $\boldsymbol{\mu}_{\theta_{ij}} = f_{\psi}(\boldsymbol{t}_i, \boldsymbol{l}_j)$
    $\boldsymbol{\Sigma}_{\theta_{ij}} = f_{\phi}(\boldsymbol{t}_i, \boldsymbol{l}_j)$
    $\boldsymbol{\vartheta}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{\theta_{ij}}, \boldsymbol{\Sigma}_{\theta_{ij}})$
    for $\mathbf{x}_{ijk} \in X_{ij}$
      $y_{ijk} \sim p(\cdot \mid \mathbf{x}_{ijk}, \boldsymbol{\vartheta}_{ij})$

pairs $\mathcal{S}$. However, as we estimate all task–language parameter vectors $\boldsymbol{\vartheta}_{ij}$ jointly, the model allows us to draw inferences about the parameters for pairs in $\mathcal{U}$ as well. The intuition for why this should work is as follows: By observing multiple pairs where the task (language) is the same but the language (task) varies, the model learns to distil the relevant knowledge for zero-shot learning because the generative model structurally enforces a disentangled representations—separating representations for the tasks from the representations for the languages rather than lumping them together into a single entangled representation (Wu and Dredze, 2019, *inter alia*). Furthermore, the neural networks f_{ψ} and f_{ϕ} mapping the task- and language-specific latent variables to neural parameters are shared allowing the model to generalise across task–language pairs.

5.3 Variational Inference

Exact computation of the posterior over the latent variables $p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l} \mid \mathbf{x})$ is intractable. Thus, we need to resort to an approximation. In this chapter, I use variational inference as an approximate inference scheme. Variational inference finds an approximate posterior over the latent variables by minimising the variational gap, which may be expressed as the Kullback–Leibler (KL) divergence between the variational approximation $q(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l})$ and the true posterior $p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l} \mid \mathbf{x})$. In this chapter, I employ the following variational distributions:

$$q_{\boldsymbol{\lambda}} = \mathcal{N}(\boldsymbol{m}_t, \boldsymbol{S}_t) \qquad \boldsymbol{m}_t \in \mathbb{R}^h, \boldsymbol{S}_t \in \mathbb{R}^{h \times h} \tag{5.1}$$

$$q_{\boldsymbol{\nu}} = \mathcal{N}(\boldsymbol{m}_l, \boldsymbol{S}_l) \qquad \boldsymbol{m}_l \in \mathbb{R}^h, \boldsymbol{S}_l \in \mathbb{R}^{h \times h} \tag{5.2}$$

$$q_{\boldsymbol{\xi}} = \mathcal{N}(f_\psi(\boldsymbol{t}, \boldsymbol{l}), \operatorname{diag}(f_\phi(\boldsymbol{t}, \boldsymbol{l}))) \tag{5.3}$$

Note the unusual choice to tie parameters between the generative model and the variational family in Equation (5.3); however, I found that this performs better in practice based on the final results of my experiments.

Through a standard algebraic manipulation in Equation (5.4), the KL-divergence for the generative model can be shown to equal the marginal log-likelihood $\log p(\mathbf{x})$, independent from $q(\cdot)$, and the so-called evidence lower bound (ELBO) $\mathcal{L}$. Thus, approximate inference becomes an optimisation problem where maximising $\mathcal{L}$ results in minimising the KL-divergence. One derives $\mathcal{L}$ is by expanding the marginal log-likelihood as in Equation (5.5) by means of Jensen's inequality. I also show that $\mathcal{L}$ can be further broken into a series of terms as illustrated in Equation (5.7). In particular, we see that it is only the first term in the expansion that requires approximation. The subsequent terms are KL-divergences between variational and true distributions that have closed-form solution due to my choice of prior. Due to the parameter-tying scheme above, the KL-divergence in Equation (5.7) between the variational distribution $q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})$ and the prior distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})$ is zero.

In general, the covariance matrices $\boldsymbol{S}_t$ and $\boldsymbol{S}_l$ in Equation (5.1) and Equation (5.2) will require $\mathcal{O}(h^2)$ space to store. As h is often very large, it is impractical to materialise either matrix in its entirety. Thus, in this chapter, I experiment with smaller matrices that have a reduced memory footprint; specifically, I consider a *diagonal* covariance matrix and a *diagonal plus low-rank* covariance structure. A diagonal covariance matrix makes computation feasible with a complexity of $\mathcal{O}(h)$; this, however, comes at the cost of failing to capture the complex interactions among parameters, since non-diagonal elements are zero. To allow for a more expressive variational family, I also consider a covariance matrix that is the sum of a diagonal matrix and a low-rank matrix:

$$\boldsymbol{S}_t = \operatorname{diag}(\boldsymbol{\delta}_t^2) + \boldsymbol{B}_t \boldsymbol{B}_t^\top \tag{5.8}$$

$$\boldsymbol{S}_l = \operatorname{diag}(\boldsymbol{\delta}_l^2) + \boldsymbol{B}_l \boldsymbol{B}_l^\top \tag{5.9}$$

where $\boldsymbol{B} \in \mathbb{R}^{h \times k}$ ensures that $\operatorname{rank}\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \leq k$, and $\operatorname{diag}(\boldsymbol{\delta})$ is diagonal. We can store this structured covariance matrix in $\mathcal{O}(kh)$ space.

$$\begin{aligned}
\mathbb{KL}\left(q(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l}) \,||\, p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l} \mid \mathbf{x})\right) &= -\mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l} \mid \mathbf{x})}{q(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l})} \\
&= -\mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \left[\log p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l}, \mathbf{x}) - \log p(\mathbf{x}) - \log q(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l})\right] \\
&= \log p(\mathbf{x}) - \mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l}, \mathbf{x})}{q(\boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l})} \triangleq \log p(\mathbf{x}) - \mathcal{L}
\end{aligned} \tag{5.4}$$

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \left(\iiint p(\mathbf{x}, \boldsymbol{\vartheta}, \boldsymbol{t}, \boldsymbol{l}) \,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\boldsymbol{l} \right) \\
&= \log \left(\iiint p(\mathbf{x} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})\, p(\boldsymbol{t})\, p(\boldsymbol{l}) \,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\boldsymbol{l} \right) \\
&= \log \left(\iiint \frac{q_{\boldsymbol{\lambda}}(\boldsymbol{t})\, q_{\boldsymbol{\nu}}(\boldsymbol{l})\, q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})}{q_{\boldsymbol{\lambda}}(\boldsymbol{t})\, q_{\boldsymbol{\nu}}(\boldsymbol{l})\, q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})} p(\mathbf{x} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})\, p(\boldsymbol{t})\, p(\boldsymbol{l}) \,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\boldsymbol{l} \right) \\
&= \log \left(\mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \frac{p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})\, p(\boldsymbol{t})\, p(\boldsymbol{l})\, p(\mathbf{x} \mid \boldsymbol{\vartheta})}{q_{\boldsymbol{\lambda}}(\boldsymbol{t})\, q_{\boldsymbol{\nu}}(\boldsymbol{l})\, q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})} \right) \\
&\geq \mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \left[\log \frac{p(\mathbf{x} \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})\, p(\boldsymbol{t})\, p(\boldsymbol{l})}{q_{\boldsymbol{\lambda}}(\boldsymbol{t})\, q_{\boldsymbol{\nu}}(\boldsymbol{l})\, q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})} \right] \triangleq \mathcal{L} && (5.5) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{t}\sim q_{\boldsymbol{\lambda}}} \mathop{\mathbb{E}}_{\boldsymbol{l}\sim q_{\boldsymbol{\nu}}} \left[\mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \left[\log p(\mathbf{x} \mid \boldsymbol{\vartheta}) + \log \frac{p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})}{q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})} \right] + \log \frac{p(\boldsymbol{t})}{q_{\boldsymbol{\lambda}}(\boldsymbol{t})} + \log \frac{p(\boldsymbol{l})}{q_{\boldsymbol{\nu}}(\boldsymbol{l})} \right] \\
&= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \log p(\mathbf{x} \mid \boldsymbol{\vartheta})}_{\textit{requires approximation}} - && (5.6) \\
&\quad - \underbrace{\left(\mathbb{KL}\left(q_{\boldsymbol{\lambda}}(\boldsymbol{t}) \,||\, p(\boldsymbol{t})\right) + \mathbb{KL}\left(q_{\boldsymbol{\nu}}(\boldsymbol{l}) \,||\, p(\boldsymbol{l})\right) + \mathbb{KL}\left(q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l}) \,||\, p(\boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{l})\right) \right)}_{\textit{closed-form solution}} && (5.7)
\end{aligned}$$

By definition, covariance matrices must be symmetric and positive semi-definite. The first property holds by construction. The second property is enforced by a softplus parameterization where $\text{softplus}(\cdot) \triangleq \ln(1+\exp(\cdot))$. Specifically, I define $\boldsymbol{\delta}^2 = \text{softplus}(\boldsymbol{\rho})$ and optimise over $\boldsymbol{\rho}$.

5.3.1 Stochastic Variational Inference

To speed up the training time, I make use of *stochastic* variational inference (Hoffman et al., 2013). In this setting, I randomly sample a task $t_i \in T$ and language $l_j \in L$

among seen combinations during each training step,[4] and randomly select a batch of examples from the dataset for the sampled task–language pair. I then optimise the parameters of the feed-forward neural networks $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ as well as the parameters of the variational approximation to the posterior $\boldsymbol{m}_t$, $\boldsymbol{m}_l$, $\boldsymbol{\rho}_t$, $\boldsymbol{\rho}_l$, $\boldsymbol{B}_t$ and $\boldsymbol{B}_l$ with a stochastic gradient-based optimiser (discussed in Section 5.4.2).

The KL divergence terms and their gradients in the ELBO appearing in Equation (5.7) can be computed in closed form as the relevant densities are Gaussian (Duchi, 2007, p. 13). Moreover, they can be calculated for Gaussians with diagonal and diagonal plus low-rank covariance structures without explicitly unfolding the full matrix. For a choice of prior $p = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a diagonal plus low-rank covariance structure, we have:

$$\mathbb{KL}\left(q \,||\, p\right) = \frac{1}{2}\Big[\sum_{i=1}^{h}(\boldsymbol{m}_i^2 + \boldsymbol{\delta}_i^2 + \sum_{j=1}^{k}\boldsymbol{B}_{ij}^2) - h - \ln\det(\boldsymbol{S})\Big] \tag{5.10}$$

where $\boldsymbol{B}_{ij}$ is the element in the i-th row and j-th column of $\boldsymbol{B}$. This derives from the general formula for computing the KL-divergence between multivariate Gaussians analytically:

$$\mathbb{KL}\left(q \,||\, p\right) = \frac{1}{2}\Big[\ln\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{S}|} - d + \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + (\boldsymbol{\mu} - \boldsymbol{m})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{m})\Big] \tag{5.11}$$

By substituting $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, it is trivial to obtain Equation (5.10).

The last term of Equation (5.10) can be estimated without computing the full matrix explicitly thanks to the generalisation of the matrix–determinant lemma,[5] which, applied to the factored covariance structure, yields:

$$\ln\det(\boldsymbol{S}) = \ln\Big[\det(\boldsymbol{I} + \boldsymbol{B}^\top\operatorname{diag}(\boldsymbol{\delta}^{-2})\boldsymbol{B})\Big] + \sum_{i=1}^{h}\ln(\boldsymbol{\delta}_i^2) \tag{5.12}$$

where $\boldsymbol{I} \in \mathbb{R}^k$. The KL divergence for the variant with diagonal covariance is just a special case of Equation (5.10) with $\boldsymbol{B}_{ij} = 0$.

[4] As an alternative, I experimented with a setup where sampling probabilities are proportional to the number of examples of each task–language combination, but this achieved similar performances on the development sets.

[5] $\det(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^\top) = \det(\boldsymbol{I} + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{U}) \cdot \det(\boldsymbol{A})$. Note that the lemma assumes that $\boldsymbol{A}$ is invertible.

However, as stated before, the following expectation does not admit a closed-form solution. Thus I consider a Monte Carlo approximation:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{\vartheta}\sim q_{\boldsymbol{\xi}}} \log p(\mathbf{x} \mid \boldsymbol{\vartheta}) = \int q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta}) \log p(\mathbf{x} \mid \boldsymbol{\vartheta}) \, \mathrm{d}\boldsymbol{\vartheta} \\ &\approx \frac{1}{V} \sum_{v=1}^{V} \log p(\mathbf{x} \mid \boldsymbol{\vartheta}^{(v)}) \quad \text{where } \boldsymbol{\vartheta}^{(v)} \sim q_{\boldsymbol{\xi}} \end{aligned} \tag{5.13}$$

where V is the number of Monte Carlo samples taken. In order to allow the gradient to easily flow through the generated samples, I adopt the re-parametrization trick (Kingma and Welling, 2014). Specifically, I exploit the following identities $\boldsymbol{t}_i = \boldsymbol{\mu}_{t_i} + \boldsymbol{\sigma}_{t_i} \odot \boldsymbol{\epsilon}$ and $\boldsymbol{l}_j = \boldsymbol{\mu}_{l_j} + \boldsymbol{\sigma}_{l_j} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\odot$ is the Hadamard product. For the diagonal plus low-rank covariance structure, I exploit the identity:

$$\boldsymbol{\mu} + \mathrm{diag}(\boldsymbol{\delta}^2 \odot \boldsymbol{\epsilon}) + \boldsymbol{B}\boldsymbol{\zeta} \tag{5.14}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^h$, $\boldsymbol{\zeta} \in \mathbb{R}^k$, and both are sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The mean $\boldsymbol{\mu}_{\theta_{ij}}$ and the diagonal of the covariance matrix $\boldsymbol{\sigma}^2_{\theta_{ij}}$ are deterministically computed given the above samples and the parameters $\boldsymbol{\vartheta}_{ij}$ are sampled from $\mathcal{N}(\boldsymbol{\mu}_{\theta_{ij}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\theta_{ij}}))$, again with the re-parametrization trick.

5.3.2 Posterior Predictive Distribution

During test time, I perform zero-shot predictions on an unseen task–language pair by plugging in the posterior means (under the variational approximation) into the model. As an alternative, I experimented with ensemble predictions through Bayesian model averaging. I.e., for data for seen combinations $\mathbf{x}_{\mathcal{S}}$ and data for unseen combinations $\mathbf{x}_{\mathcal{U}}$, the true predictive posterior can be approximated as $p(\mathbf{x}_{\mathcal{U}} \mid \mathbf{x}_{\mathcal{S}}) = \int p(\mathbf{x}_{\mathcal{U}} \mid \boldsymbol{\vartheta}, \mathbf{x}_{\mathcal{S}}) \, q_{\boldsymbol{\xi}}(\boldsymbol{\vartheta} \mid \mathbf{x}_{\mathcal{S}}) \, \mathrm{d}\boldsymbol{\vartheta} \approx \frac{1}{V} \sum_{v=1}^{V} p(\mathbf{x}_{\mathcal{U}} \mid \boldsymbol{\vartheta}^{(v)}, \mathbf{x}_{\mathcal{S}})$, where V are 100 Monte Carlo samples from the posterior $q_{\boldsymbol{\xi}}$. Performances on the development sets are comparable to simply plugging in the posterior mean. However, foreshadowing Section 5.5.4, marginalisation during prediction has additional advantages, such as better estimating uncertainty.

5.4 Experimental Setup

5.4.1 Data

I select NER and POS tagging as experimental tasks because their datasets encompass an ample and diverse sample of languages, and are common benchmarks for resource-poor NLP (Cotterell and Duh, 2017, *inter alia*). In particular, I opt for WikiANN (Pan et al., 2017) for the NER task and Universal Dependencies 2.4 (UD; Nivre et al., 2019) for POS tagging. Our sample of languages is chosen from the intersection of those available in WikiANN and UD. However, I remark that this sample is heavily biased towards the Indo-European family (Gerz et al., 2018b). Instead, the selection should be: i) typologically diverse, to ensure that the evaluation scores truly reflect the expected cross-lingual performance (Ponti et al., 2020); ii) a mixture of resource-rich and low-resource languages, to recreate a realistic setting and to allow for studying the effect of data size. Hence, I further filter the languages in order to make the sample more balanced. In particular, I sub-sample Indo-European languages by including only resource-poor ones, and keep all the languages from other families. Our final sample comprises 33 languages from 4 continents (17 from Asia, 11 from Europe, 4 from Africa, and 1 from South America) and from 11 families (6 Uralic, 6 Indo-European, 5 Afroasiatic, 3 Niger-Congo, 3 Turkic, 2 Austronesian, 2 Dravidian, 1 Austroasiatic, 1 Kra-Dai, 1 Tupian, 1 Sino-Tibetan), as well as 2 isolates. The full list of language ISO 639-2 codes is reported in Figure 5.2.

In order to simulate a zero-shot setting, I hold out in turn half of all possible task–language pairs and regard them as unseen, while treating the others as seen pairs. The partition is performed in such a way that a held-out pair has data available for the same task in a different language, and for the same language in a different task.[6] Under this constraint, pairs are assigned to train or evaluation at random.[7]

I randomly split the WikiANN datasets into training, development, and test portions with a proportion of 80-10-10. I use the provided splits for UD; if the training set for a language is missing, I treat the test set as such when the language is held out, and as a training set when it is among the seen pairs.[8]

[6] I use the controlled partitioning for the following reason. If a language lacks data both for NER and for POS, the proposed factorisation method cannot provide estimates for its posterior. I leave model extensions that can handle such cases for future work.

[7] See Section 5.5.2 for further experiments on splits controlled for language distance and sample size.

[8] Note that, in the second case, no evaluation takes place on such language.

5.4.2 Hyper-parameters

The multilingual M-BERT encoder is initialised with parameters pre-trained on masked language modelling and next sentence prediction on 104 languages (Devlin et al., 2019).[9] I opt for the cased BERT-BASE architecture, which consists of 12 layers with 12 attention heads and a hidden size of 768. As a consequence, this is also the dimension e of each encoded WordPiece unit, a subword unit obtained through BPE (Wu et al., 2016). The dimension h of the multivariate Gaussian for task and language latent variables is set to 100. The deep feed-forward networks f_ψ and f_ϕ have 6 layers with a hidden size of 400 for the first layer, 768 for the internal layers, and ReLU non-linear activations. Their depth and width were selected based on validation performance.

The expectations over latent variables in Equation (5.7) are approximated through 3 Monte Carlo samples per batch during training. The KL terms are weighted with $\frac{1}{|K|}$ uniformly across training, where $|K|$ is the number of mini-batches.[10] All the means $\boldsymbol{m}$ of the variational approximation are initialised with a random sample from $\mathcal{N}(0, 0.1)$, and the parameters for covariance matrices $\boldsymbol{S}$ with a random sample from $\mathcal{U}(0, 0.5)$, following Stolee and Patterson (2019). $k = 10$ is chosen as the number of columns of $\boldsymbol{B}$ so it fits into memory. The maximum sequence length for inputs is limited to 250. The batch size is set to 8, and the best setting for the Adam optimiser (Kingma and Ba, 2015) was found to be an initial learning rate of $5 \cdot 10^{-6}$ based on grid search. In order to avoid over-fitting, I perform early stopping with a patience of 10 and a validation frequency of 2.5K steps.

5.4.3 Baselines

I consider four baselines for cross-lingual transfer that also use BERT as an encoder shared across all languages.

First Baseline. A common approach is transfer from the **nearest source** (NS) language, which selects the most compatible source to a target language in terms of similarity. In particular, the selection can be based on family membership (Cotterell and Heigold, 2017; Kann et al., 2017; Zeman and Resnik, 2008), typological features (Deri and Knight, 2016), KL-divergence between part-of-speech trigram distributions (Agić, 2017; Rosa and Zabokrtsky, 2015), tree edit distance of delexicalized dependency parses (Ponti et al., 2018a), or a combination of the above (Lin et al., 2019). In this chapter, during

[9] Available at github.com/google-research/bert/blob/master/multilingual.md

[10] I found this weighting strategy to work better than annealing as proposed by Blundell et al. (2015).

evaluation, I choose the classifier associated with the observed language with the highest cosine similarity between its typological features and those of the held-out language. These features are sourced from URIEL (Littell et al., 2017) and contain information about family, area, syntax, and phonology.

Second Baseline. I also consider transfer from the **largest source** (LS) language, i.e. the language with most training examples. This approach has been adopted by several recent works on cross-lingual transfer (Artetxe et al., 2020; Conneau et al., 2018, *inter alia*). In my implementation, I always select the English classifier for prediction.[11] In order to make this baseline comparable to my model, I adjust the number of English NER training examples to the sum of the examples available for all seen languages $\mathcal{S}$.[12]

Third Baseline. Next, I apply a protocol designed by Rahimi et al. (2019) for weighting the predictions of a classifier ensemble according to their reliability. For a specific task, the reliability of each language-specific classifier is estimated through a Bayesian graphical model. Intuitively, this model learns from error patterns, which behave more randomly for untrustworthy models and more consistently for the others. Among the protocols proposed in the paper, I opt for **BEA** in its zero-shot, token-based version, as it achieves the highest scores in a setting comparable to the current experiment. I refer the reader to the original paper for the details.[13]

Fourth Baseline. Finally, I take inspiration from Wu and Dredze (2019). The **joint multilingual** (JM) baseline, contrary to the previous baselines, consists of two classifiers (one for POS tagging and another for NER) shared among all observed languages for a specific task. I follow the original implementation of Wu and Dredze (2019) closely adopting all recommended hyper-parameters and strategies, such as freezing the parameters of all encoder layers below the 3$^{\text{rd}}$ for sequence labelling tasks.

It must be noted that the number of parameters in the generative model scales better than baselines with language-specific classifiers, but worse than those with language-agnostic classifiers, as the number of languages grows. However, even in the second case, increasing the depth of baselines networks to match the parameter count is detrimental if the BERT encoder is kept trainable, which was also verified in previous work (Peters et al., 2019).

[11] I include English to make the baseline more competitive, but note that this language is not available for the generative model as it is both Indo-European and resource-rich.

[12] The number of NER training examples is 1,093,184 for the first partition and 520,616 for the second partition.

[13] I implemented this model through the original code at github.com/afshinrahimi/mmner.

Task	BEA	NS	LS	JM	PF-d	PF-lr
POS	47.65±1.54	42.84±1.23	60.51±0.43	64.04±0.18	**65.00±0.12**	**64.71±0.18**
NER	66.45±0.56	74.16±0.56	78.97±0.56	85.65±0.13	86.26±0.17	**86.70±0.10**

Table 5.1 Results per task averaged across all languages.

5.5 Results and Discussion

5.5.1 Zero-shot Transfer

Firstly, I present the results for zero-shot prediction based on the generative model using both of the approximate inference schemes (with diagonal covariance **PF-d** and factor covariance **PF-lr**). Table 5.1 summarises the results on the two tasks of POS tagging and NER averaged across all languages. Our model (in both its variants) outperforms the four baselines on both tasks, including state-of-the-art alternative methods. In particular, PF-d and PF-lr gain 4.49 / 4.20 in accuracy (~7%) for POS tagging and 7.29 / 7.73 in F1 score (~10%) for NER on average compared to transfer from the largest source (**LS**), the strongest baseline for single-source transfer. Compared to multilingual joint transfer from multiple sources (**JM**), the two variants gain 0.95 / 0.67 in accuracy (~1%) for POS tagging and +0.61 / +1.05 in F1 score (~1%).

More details about the individual results on each task–language pair are provided in Figure 5.2, which includes the mean of the results over 3 separate runs. Overall, I obtain improvements in 23/33 languages for NER and on 27/45 treebanks for POS tagging, which further supports the benefits of transferring both from tasks and languages.

Considering the baselines, the relative performance of LS versus NS is an interesting finding *per se*. LS largely outperforms NS on both POS tagging and NER. This shows that having more data is more informative than relying primarily on similarity according to linguistic properties. This finding contradicts the received wisdom (Cotterell and Heigold, 2017; Lin et al., 2019; Rosa and Zabokrtsky, 2015, *inter alia*) that related languages tend to be the most reliable source. I conjecture that this is due to the pre-trained multi-lingual BERT encoder, which helps to bridge the gap between unrelated languages (Wu and Dredze, 2019).

The two baselines that hinge upon transfer from multiple sources lie on opposite sides of the spectrum in terms of performance. On the one hand, BEA achieves the lowest average score for NER, and surpasses only NS for POS tagging. I speculate that this is

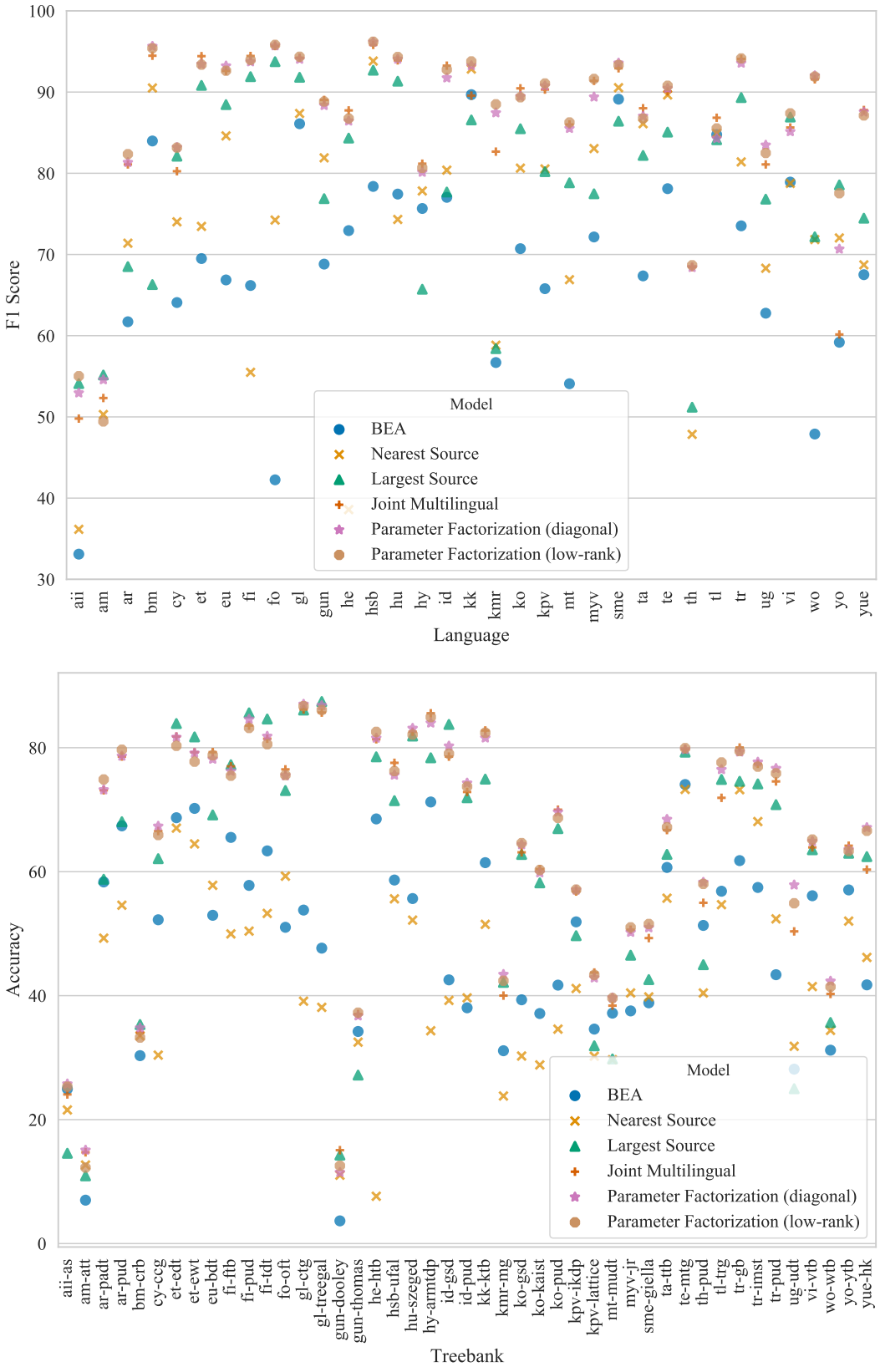


Figure 5.2 Results for NER (top) and POS tagging (bottom): four baselines for cross-lingual transfer compared to Matrix Factorisation with diagonal covariance and diagonal plus low-rank covariance.

Task	$\|L\| = 11$		$\|L\| = 22$	
	Sim	Dif	Sim	Dif
POS	72.44	53.25	66.59	63.22
NER	89.51	81.73	86.78	85.12

Table 5.2 Average performance when relying on $|L|$ similar (*Sim*) versus different (*Dif*) languages in the train and evaluation sets.

due to the following: i) adapting the protocol from Rahimi et al. (2019) to my model implies assigning a separate classifier head to each task–language pair, each of which is exposed to fewer examples compared to a shared one. This fragmentation fails to take advantage of the massively multilingual nature of the encoder; ii) my language sample is more typologically diverse, which means that most source languages are unreliable predictors. On the other hand, JM yields extremely competitive scores. Similarly to my model, it integrates knowledge from multiple languages and tasks. The extra boost in my model stems from its ability to disentangle each aspect of such knowledge and recombine it appropriately.

Moreover, comparing the two approximate inference schemes from Section 5.3.1, PF-lr obtains a small but statistically significant improvement over PF-d in NER, whereas they achieve the same performance on POS tagging. This means that the posterior is modelled well enough by a Gaussian where covariance among co-variates is negligible.

We can see that even for the best model (PF-lr) there is a wide variation in the scores for the same task across languages. POS tagging accuracy ranges from 12.56 ± 4.07 in Guaraní to 86.71 ± 0.67 in Galician, and NER F1 scores range from 49.44 ± 0.69 in Amharic to 96.20 ± 0.11 in Upper Sorbian. Part of this variation is explained by the fact that the multilingual BERT encoder is not pre-trained in a subset of these languages (e.g., Amharic, Guaraní, Uyghur). Another cause is more straightforward: the scores are expected to be lower in languages for which we have fewer training examples in the seen task–language pairs.

5.5.2 Language Distance and Sample Size

While I designed the language sample to be both realistic and representative of the cross-lingual variation, there are several factors inherent to a sample that can affect the zero-shot transfer performance: i) *language distance*, the similarity between seen and held-out languages; and ii) *sample size*, the number of seen languages. In order to

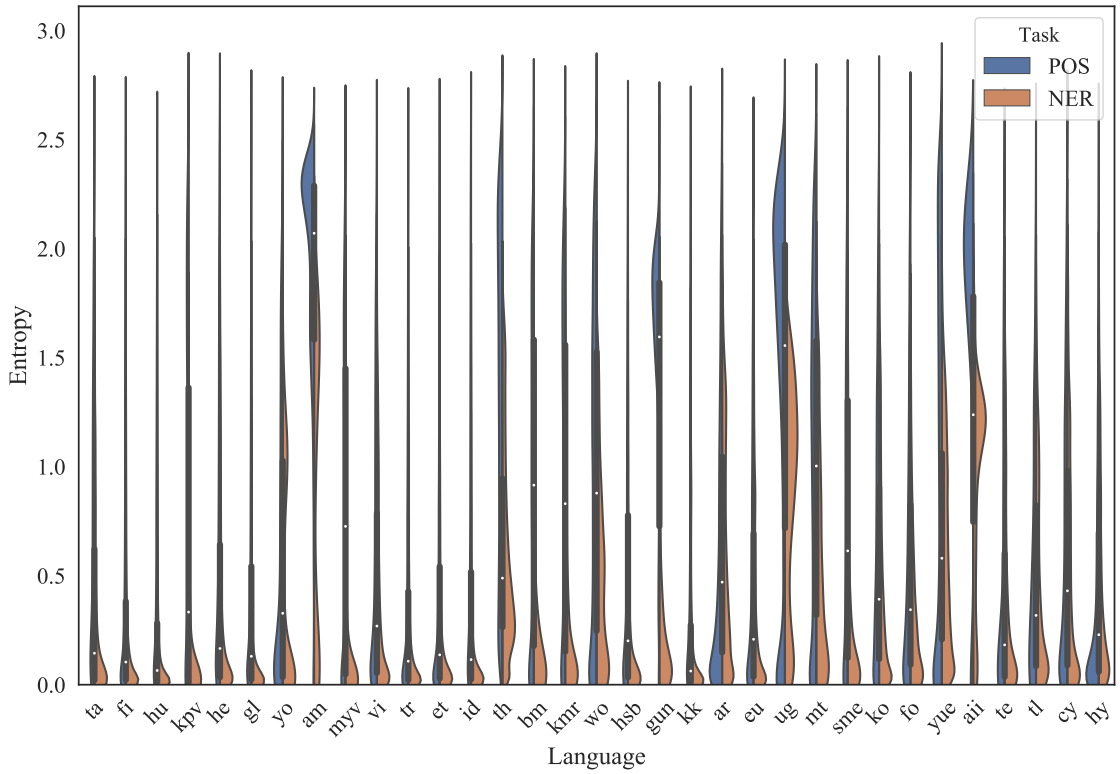


Figure 5.3 Entropy of the posterior predictive distributions over classes for each test example. The higher the entropy, the more uncertain the prediction.

disentangle these factors, I construct subsets of size $|L|$ so that training and evaluation languages are either maximally similar (*Sim*) or maximally different (*Dif*). As a proxy measure, I consider as 'similar' languages belonging to the same family. In Table 5.2, I report the performance of parameter factorisation with diagonal plus low-rank covariance (PF-lr), the best model from Section 5.5.1, for each of these subsets.

Based on Table 5.2, there emerges a trade-off between language distance and sample size. In particular, performance is higher in *Sim* subsets compared to *Dif* subsets for both tasks (POS and NER) and for both sample sizes $|L| \in \{11, 22\}$. In larger sample sizes, the average performance increases for *Dif* but decreases for *Sim*. Intuitively, languages with labelled data for several relatives benefit from small, homogeneous subsets. Introducing further languages introduces noise. Instead, languages where this is not possible (such as isolates) benefit from an increase in sample size.

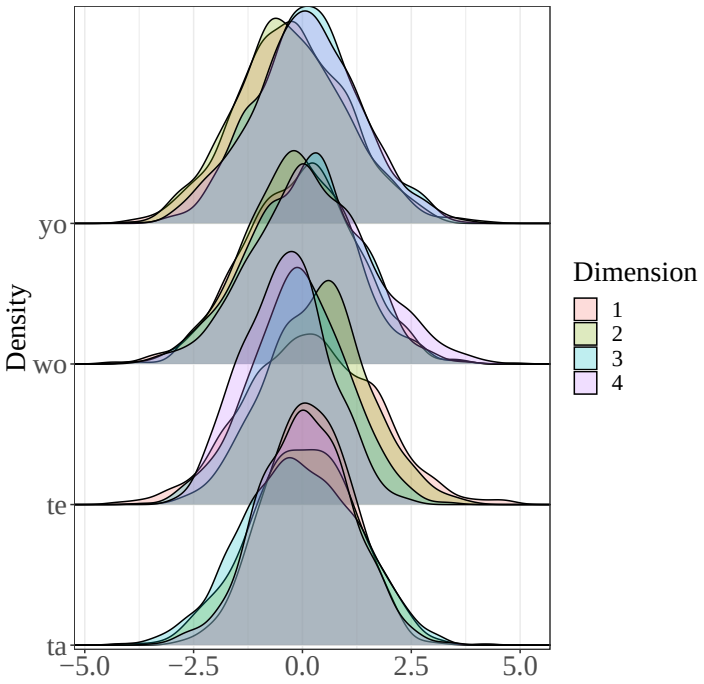


Figure 5.4 Samples from the posteriors of 4 languages, PCA-reduced to 4 dimensions.

5.5.3 Visualisation of the Learned Posteriors

The approximate posteriors of the latent variables can be visualised in order to study the learned representations for languages. Previous work (Bjerva and Augenstein, 2018; Johnson et al., 2017; Malaviya et al., 2017; Östling and Tiedemann, 2017) induced point estimates of language representations from artificial tokens concatenated to every input sentence, or from the aggregated values of the hidden state of a neural encoder. The information contained in such representations depends on the task (Bjerva and Augenstein, 2018), but mainly reflects the structural properties of each language (Bjerva et al., 2019).

In our work, due to the estimation procedure, languages are represented by full distributions rather than point estimates. By inspecting the learned representations, language similarities do not appear to follow the structural properties of languages. This is most likely due to the fact that parameter factorisation takes place *after* the multilingual BERT encoding, which blends the structural differences across languages. A fair comparison with previous works without such an encoder is left for future investigation.

As an example, consider two pairs of languages from two distinct families: Yoruba and Wolof are Niger-Congo from the Atlantic-Congo branch, Tamil and Telugu are Dravidian. We take 1,000 samples from the approximate posterior over the latent variables for each of these languages. In particular, we focus on the variational scheme with a low-rank covariance structure. We then reduce the dimensionality of each sample to 4 through PCA,[14] and we plot the density along each resulting dimension in Figure 5.4. We observe that density areas of each dimension do not necessarily overlap between members of the same family. Hence, the learned representations depend on more than genealogy.

5.5.4 Entropy of the Predictive Distribution

A notable problem of point estimate methods is their tendency to assign most of the probability mass to a single class even in scenarios with high uncertainty. Zero-shot transfer is one of such scenarios, because it involves drastic distribution shifts in the data (Rabanser et al., 2019). A key advantage of Bayesian inference, instead, is marginalisation over parameters, which yields smoother posterior predictive distributions (Kendall and Gal, 2017; Wilson, 2019).

I run an analysis of predictions based on (approximate) Bayesian model averaging. First, I randomly sample 800 examples from each test set of a task–language pair. For each example, I predict a distribution over classes Y through model averaging based on 10 samples from the posteriors. I then measure the prediction entropy of each example, i.e. $\mathrm{H}(p) = -\sum_{y}^{|Y|} p(Y = y) \ln p(Y = y)$, whose plot is shown in Figure 5.3.

Entropy is a measure of uncertainty. Intuitively, the uniform categorical distribution (maximum uncertainty) has the highest entropy, whereas if the whole probability mass falls into a single class (maximum confidence), then the entropy $\mathrm{H} = 0$.[15] As it emerges from Figure 5.3, predictions in certain languages tend to have higher entropy on average, such as in Amharic, Guaraní, Uyghur, or Assyrian Neo-Aramaic. This aligns well with the performance metrics in Figure 5.2. In practice, languages with low scores tend to display high entropy in the predictive distribution, as expected.

To verify this claim, I measure the Pearson's correlation between entropies of each task–language pair in Figure 5.3 and the log-likelihood, a performance metric. I find a very strong negative correlation with a coefficient of $\rho = -0.914$ and a two-tailed p-value of 1.018×10^{-26}. Finally, I compare this inference scheme (based on a stochastic variational approximation) with point estimate methods. In particular, I measure the same correlation between predictive entropy and performance for Monte Carlo Dropout

[14] Note that the dimensionality reduced samples are also Gaussian since PCA is a linear method.
[15] The maximum entropy is ≈ 2.2 for 9 classes as in NER and ≈ 2.83 for 17 classes as in POS tagging.

(MCD; Gal and Ghahramani, 2016a) under the same model. MCD is an ensemble method where different parameter values are sampled simply by applying random dropout patterns to the same maximum-likelihood estimate. I found, again, a negative correlation, but with a smaller coefficient of $\rho = -0.634$. From this result, we may conclude that SVI inference better characterises predictive uncertainty.

5.6 Related Work

Our approach builds on ideas from several different fields: cross-lingual transfer in NLP, with a particular focus on matrix factorisation, contextual parameter generation, and neural Bayesian methods.

Data Matrix Factorisation. Although I am the first to propose a factorisation of the *parameter* space for unseen combinations of tasks and languages, the factorisation of *data* for collaborative filtering and social recommendation is an established research area. In particular, the missing values in sparse data structures such as user-movie review matrices can be filled via probabilistic matrix factorisation (PMF) through a linear combination of user and movie matrices (Ma et al., 2008; Mnih and Salakhutdinov, 2008; Shan and Banerjee, 2010, *inter alia*) or through neural networks (Dziugaite and Roy, 2015). Inference for PMF can be carried out through MAP inference (Dziugaite and Roy, 2015), Markov chain Monte Carlo (MCMC; Salakhutdinov and Mnih, 2008) or stochastic variational inference (Stolee and Patterson, 2019). Contrary to prior work, I perform factorisation on latent variables (task- and language-specific parameters) rather than observed ones (data).

Contextual Parameter Generation. Our model is reminiscent of the idea that parameters can be conditioned on language representations, as proposed by Platanios et al. (2018). However, since this approach is limited to a single task and a joint learning setting, it is not suitable for generalisation in a zero-shot transfer setting.

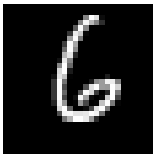
Bayesian Neural Networks. So far, these models have found only limited application in NLP for resource-poor languages, despite their desirable properties. Firstly, they can incorporate priors over parameters to endow neural networks with the correct inductive biases towards language: Ponti et al. (2019b) constructed a prior imbued with universal linguistic knowledge for zero- and few-shot character-level language modelling. Secondly, they avoid the risk of over-fitting by taking into account uncertainty. For instance,

Shareghi et al. (2019) and Doitch et al. (2019) use a perturbation model to sample high-quality and diverse solutions for structured prediction in cross-lingual parsing.

5.7 Conclusion

The main contribution of this chapter is a Bayesian generative model for multiple NLP tasks and languages. At its core lies the idea that the space of neural weights can be factorised into latent variables for each task and each language. While training data are available only for a meagre subset of task–language combinations, this model opens up the possibility to perform prediction in novel, undocumented combinations at evaluation time. I performed inference through stochastic variational methods, and ran experiments on zero-shot named entity recognition (NER) and part-of-speech (POS) tagging in a typologically diverse set of 33 languages. Based on the reported results, I conclude that leveraging the information from tasks and languages simultaneously is superior to model transfer from English (relying on more abundant in-task data in the source language), from the most typologically similar language (relying on prior information on language relatedness), or from multiple source languages. Moreover, I found that the entropy of predictive posterior distributions obtained through Bayesian model averaging correlates almost perfectly with the error rate in the prediction. As a consequence, my approach holds promise to alleviating data paucity issues for a wide spectrum of languages and tasks, and to make knowledge transfer more robust to uncertainty.

Finally, I remark that my model is amenable to be extended to multilingual tasks beyond sequence labelling—such as natural language inference (Conneau et al., 2018) and question answering (Artetxe et al., 2020; Clark et al., 2020; Lewis et al., 2019)—and to zero-shot transfer across combinations of multiple modalities (e.g. speech, text, and vision) with tasks and languages. I leave these exciting research threads for future research.

Chapter 

Conclusions

This final chapter draws some main conclusions from the research carried out in this thesis. In Section 6.1, I provide a synopsis of the motivations justifying the endeavour of conforming machine learning to some aspects of human learning, including the presence of an inductive bias accelerating learning and the capacity to generalise by disentangling and recombining knowledge. Afterwards, I assess to what extent the proposed Bayesian neural framework satisfied these desiderata, and in particular sample efficiency, resilience to catastrophic forgetting, generalisation to novel domains, and robustness to uncertainty. I take stock of the findings and contributions of this thesis in this respect in Section 6.2, and discuss the implications in Section 6.3. Finally, I speculate about possible perspectives for future work and address the remaining open questions in Section 6.4.

6.1 Motivation Synopsis

The present thesis set out from the ostensible disconnect between language acquisition in humans on the one hand, where limited examples are sufficient to master any of the multifarious language varieties the world is studded with, and current machine learning practices on the other, which are predicated on the availability of massive amounts of data (van Schijndel et al., 2019) and i.i.d. domains during both training and evaluation (Linzen, 2020). In fact, children exhibit much higher flexibility and efficiency, which can be better understood by considering evidence from their learning process. Error patterns (also known as 'emergent categories') and the preference for specific kinds of meaning-to-form mappings both point towards the presence of an inductive bias that guides learning (Bowerman, 2011; Clark, 2001; Slobin, 1973). Such inductive bias results

both from embodiment, the perceptual and cognitive constraints imposed by the human brain, and from grounding, the shared human experience of reality.

Bridging the hiatus between these two different paradigms of learning would have significant ramifications in practice. In fact, sample-efficient learning is indispensable to develop natural language processing applications that are genuinely multilingual. Indeed, most of the world's languages suffer from data paucity, because annotating data is expensive and time-consuming, and even unlabelled texts are often scarce due to the imbalance in usage of and access to the digital sphere across communities of speakers (Ponti et al., 2019a). Even the datasets with the amplest coverage—such as Wikipedia or Universal Dependencies—span across but a minute fraction of the total of existing languages. As a consequence, modelling resource-poor languages hinges upon the ability to cope with zero-shot and few-shot learning scenarios, which is notoriously challenging (Bottou and Bousquet, 2008; Ravi and Larochelle, 2017; Vinyals et al., 2016).[1] In particular, this requires both to adapt to novel information quickly and to access previously acquired relevant knowledge and recombine it in original ways in order to address unseen combinations of tasks, languages, and modalities. In other words, both sample efficiency and generalisation through modular design—intended as a mechanism to disentangle separate facets of linguistic knowledge—are necessary to deal with a diverse set of scarcely documented languages.

While recent efforts of the community concentrated on knowledge transfer to mitigate these problems, the solutions they offer are inconclusive. In fact, current techniques are based on pretraining deep Transformer-based encoders on language modelling in an unsupervised fashion and subsequently fine-tuning them on downstream tasks (Conneau et al., 2020; Wu and Dredze, 2019, *inter alia*) through many labelled examples of source resource-rich languages (Conneau et al., 2020; Wu and Dredze, 2019) and possibly few examples in a target resource-poor language (Lauscher et al., 2020). This is still insufficient for achieving satisfactory performance in few-shot learning, for several reasons: 1) it remains data-demanding for both pre-training and fine-tuning; 2) it tends to incur catastrophic forgetting as most techniques (including meta-learning) provide only points of initialisation for neural weights, failing to model variance and architecture parameters; 3) it is prone to error when applied to a radically different domain where the label distribution is shifted or the input language is different; 4) finally, it makes it hard to gauge the confidence of predictions, thus making models less robust.

[1] Recent attempts to address this problem, such as GPT-3 (Brown et al., 2020), managed to reduce the requirements of labelled data to a bare minimum, but at the cost of tremendously inflating the requirements of raw texts. Hence, they remain nonviable for resource-poor languages.

In this thesis, I argued that the desirable properties of a learning agent can instead be fulfilled in a unified Bayesian framework. In particular, I constructed a prior encompassing both neural weights and architectures, and capable of drawing information from both other languages and typological features, by inferring an approximate posterior through Laplace or variational methods. Performing Bayesian update rather than pretraining and fine-tuning also allows for preserving previous knowledge while acquiring a new one. Moreover, graphical models specify the dependence assumptions among the variables involved in an experiment. I took advantage of this to disentangle separate facets of knowledge relevant for any combination of task, language, and modality, which facilitates generalisation. Finally, I revealed (approximate) model averaging in estimating the predictive distribution to be a powerful tool to measure the confidence of a prediction.

6.2 Findings and Contributions

Given the motivations in Section 6.1, I ran a series of experiments whose main findings and contributions are listed below as bullet points, ordered by chapter of appearance.

6.2.1 A Prior over Weights for Language Modelling

In order to provide models with the correct inductive bias towards a new language, in Chapter 3 I proposed to harness two sources of information, namely texts in other languages and hand-crafted features from typological databases. I focused on the task of character-level, open-vocabulary language modelling in a typologically diverse sample of 77 languages, and I compared different priors over neural weights and different scenarios of data paucity.

- I leveraged the Laplace approximation for posterior inference over neural weights, developed by MacKay (1992) and recently shown by Kirkpatrick et al. (2017) to alleviate catastrophic forgetting, for cross-lingual transfer for the first time. This approach surpassed state-of-the-art alternative methods for transfer such as fine-tuning and uninformed priors by a large margin. Moreover, the results revealed that the abundance of data (copious out-of-domain texts) is more helpful than quality (few in-domain data points). In fact, the performances for zero-shot learning with a universal prior obtained through the proposed method are superior to few-shot learning with an uninformed prior.

- I complemented such prior by providing side information from typological features about target languages. More specifically, I conditioned neural hidden states on

them by adapting methods developed for language vector learning. In particular, feature concatenation (Östling and Tiedemann, 2017) turned out to surpass both hyper-networks[2] (Platanios et al., 2018) and baselines without typological features in a few-shot learning setting. On the other hand, in the zero-shot setting there appeared to be no benefit in adding typological features. Hence, the evidence for their usefulness is mixed. Possibly, this stems from issues of granularity, inconsistency, and neglect of intra-language variation inherent to typological databases (see Section 2.1).

- I studied the dynamics of the adaptation of the language universal prior to the specifics of a target language after observing a few examples. A strong correlation was found between zero-shot performance and similarity of unigram character between source and target languages. However, this correlation vanishes in the case of few-shot performance. This implies that the adaptation takes place quickly. Moreover, probing the learned posterior unravelled that there are clearly distinguished sets of parameters with high and low signal-to-noise ratio. Retaining the former unchanged prevents catastrophic forgetting. Finally, a quantitative analysis demonstrated that the universal prior displays cross-lingual tendencies in terms of syllable structure and consonant clusters, thus being truly imbued with universal linguistic knowledge about phonotactics.

6.2.2 A Prior over Architectures for Language Understanding

In Chapter 4, I investigated how to construct a prior over neural weights *and* architectures to facilitate natural language understanding in new languages, by building on recent developments in neural architecture search. I evaluated the model on a newly created evaluation benchmark for commonsense reasoning in 11 languages.

- I proposed a generalisation of current differentiable Neural Architecture Search methods, such as DARTS (Liu et al., 2019) and SNAS (Xie et al., 2019), which clarifies the implicit graphical model underlying them and enables more expressive inference schemes. In particular, the bi-level optimisation at the core of those methods can be interpreted as a version of empirical Bayes where a truncated approximation of the neural weights in an inner loop is the starting point for a point estimate of the neural architecture in an outer loop. Optimisation alternates between the nested loops until convergence. Under this light, there emerges a

[2]Hyper-networks are trainable functions that generate the parameters of a subordinate network.

hierarchical dependence of the two variables, neural weights and architecture parameters. What is more, whereas Neural Architecture Search has been deployed mostly in the visual domain, it is currently under-explored in the language domain. In particular, its usage has been somewhat limited to the language modelling task. In this thesis, its usage has been shown to be more broadly beneficial for downstream applications such as natural language understanding. Another contribution of this chapter consists in adapting NAS to the Transformer architecture, whereas previous work focused entirely on recurrent architectures. Most crucially, this generalisation allows for experimenting with alternative inference schemes (such as variational methods) and with different parametrisations of the variables (in my case, modelling the conditional probability of the neural weights given the architecture through a hyper-network).

- To provide a challenging multilingual benchmark as a test-bed for the proposed approach, I devised a novel dataset for cross-lingual commonsense causal reasoning in 11 languages, XCOPA, whose translations and annotations were crowd-sourced. The task is formulated as multiple-choice question answering: given a premise and a question, the machine has to select the more reasonable between two hypothetical answers. I proposed explicit metrics to quantify the typological, geographical, and family diversity of a dataset. These drove the selection of languages in the sample, according to the principle of variety maximisation. This ensures that evaluation reflects the true expected performance of a model cross-lingually and its robustness to rare features and distant languages. Finally, I streamlined an annotation protocol that hybridises translation and example adaptation. In particular, I individuated strategies to mitigate the sources of cross-lingual divergence, both cultural and grammatical, to keep examples both comparable and idiomatic.

- I benchmarked a series of state-of-art multilingual encoders on XCOPA, where XLM Large emerged victorious (Conneau et al., 2020). Moreover, I compared several transfer learning setups: based on the results, validation on target language data (rather than English) appeared superior (although by a little margin). Also, it was beneficial to increment the training set with out-of-domain but related English datasets such as SocialIQA (Sap et al., 2019). Most crucially, Neural Architecture Search based on a hierarchical Bayesian model surpassed all equivalent baselines with a fixed classifier architecture during fine-tuning, both in the zero-shot and few-shot learning settings. This demonstrates the viability and efficiency of the

proposed method, and its ability to prime the model towards the knowledge required to solve complex causal reasoning.

6.2.3 Modular Design via Parameter Factorisation

Finally, Chapter 5 pursued the idea that models should be able to disentangle the aspects of knowledge relevant to solve a specific combination of tasks and languages. This way, they can be recombined in novel ways when facing unprecedented combinations. In fact, while some of such combinations are documented with annotated data, most are not, which precludes supervised learning. However, the missing data can be compensated for by performing the high-level combinatorial generalisations typical of humans.

- I advocated for considering the space of neural parameters as a structured space, where each possible combination of task and language defines a separate cell. Accordingly, I defined a Bayesian generative model of neural parameters, where each task–language-specific set of parameters is conditioned on variables representing the task and language for a specific example. I also explored several approximate inference schemes for the posteriors of task and language Gaussian distributions: a diagonal approximation of their covariance, and a low-rank factored approximation.

- Evaluating the proposed model on zero-shot learning for 2 tasks (part-of-speech tagging and named entity recognition) and 33 typologically diverse languages, I compared its performance with state-of-the-art baselines relying on regular fine-tuning. In this case, source languages were selected according to the similarity of their typological features or the abundance of their annotated data. Based on the results, parameter factorisation yields large gains due to its ability to take advantage of jointly transferring knowledge from all source languages *and* tasks. Rather than conflating disparate facets of linguistic knowledge into a fully shared set of parameters, the proposed approach distils the regularities of each language and task into a dedicated representation.

- By virtue of variational inference, I put forth a method to obtain an estimate of the uncertainty of predictions—a crucial asset in radical domain shifts such as cross-lingual transfer—with a simple approximation of Bayesian model averaging. Crucially, I found an extremely strong correlation between the entropy of the predictive posterior and the accuracy of the model. Thus, predictions in low-confidence combinations can be rejected in block if they surpass a certain threshold. This increases the model robustness in zero-shot learning.

6.3 Implications and Discussion

The findings summarised in Section 6.2 have important implications for the two notions at the core of the present thesis, namely inductive bias and modular design. In what follows, I discuss them briefly in the context of the received wisdom in the literature.

6.3.1 Inductive Bias

The experiments in Chapter 3 and Chapter 4 elaborate on the established notion of inductive bias. In fact, this is generally interpreted as the set of assumptions that enable a model to generalise beyond samples encountered during training (Mitchell, 1980). This definition can be operationalised more formally in terms of model support (MacKay, 2003; Wilson, 2019) as follows. Given a model $\mathcal{M}$ (for instance, a neural architecture parameterised by $\boldsymbol{\alpha}$), weight parameters $\boldsymbol{\vartheta}$, and a series of datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$, the marginal likelihood of the i-th dataset equals $p(\mathcal{D}_i \mid \mathcal{M}) = \int p(\mathcal{D}_i \mid \mathcal{M}, \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta})\, \mathrm{d}\boldsymbol{\vartheta}$. The support is then defined as $\{\mathcal{D}_i \mid p(\mathcal{D}_i \mid \mathcal{M}) > 0\}$, the subset of datasets with a positive marginal likelihood. A model with the correct inductive bias for a specific dataset is consequently a model with the adequate distribution of marginal likelihood among supported datasets, peaked around the one of interest. Or, otherwise stated, a model assigning high probability to such a dataset notwithstanding the set of weight parameters chosen.

In the case of natural language processing, let us assume that each 'dataset' consists in a conceivable language with a bundle of typological features, including unlikely and impossible ones. The latter may involve phonotactics (e.g. consonant clusters that are unpronounceable given human anatomy), syntax (e.g. words with fixed linear positions in a sentence) (Moro and Chomsky, 2015), or semantics (e.g. nonsense meanings like '*Colorless green ideas sleep furiously*') (Chomsky, 1956, p. 116). A model with the correct inductive bias should have no support for such datasets. Thus, the support should stretch as far as a language is possible, in order to be able to learn any of the languages spoken around the world. In addition, the marginal likelihood should also peak around likely configurations, so that a model is facilitated learning them by contracting its parameters around the true solution efficiently.

In this thesis, I stressed how the importance of an inductive bias towards language not only stems from enhancing sample efficiency in machine learning, but also from shedding light on what sorts of linguistic features are favoured in language acquisition. In other words, *probing the content* of the inductive bias enables the extraction of positive scientific knowledge about the learning process. In Chapter 3, I showed that probing the

posterior over weights of a character-level language model yields common cross-lingual patterns in terms of syllabic structure and phonotactics. Recently, McCoy et al. (2020) reached a similar conclusion for morphological inflection. In their study, they show that a point estimate of neural weights learned through meta-learning reflects specific inductive biases in phonotactics depending on which languages the model is exposed to. For instance, by carefully selecting the training languages, a preference for syllables ending in a vowel (in Optimality Theory terms, NoCoda, cf. Smolensky and Prince 1993).

This line of research also opens up new frontiers for *measuring the quality* of inductive biases. Throughout this thesis, I focused on comparing the performance of identical models with and without the proposed inductive bias in downstream tasks under the same regimes of data paucity, i.e. zero-shot and few-shot learning. However, the advantage brought by different inductive biases could be quantified also as the difference in the number of training examples needed to reach the same level of performance. Overall, it should be noted that both these metrics are prone to flaws because, for instance, they are accentuated based on the affinity between source and target languages. I investigated this effect through similarity in unigram character distribution in Chapter 3 and family relationship in Chapter 5.

6.3.2 Modular Design

The notion of modular design presented in this thesis hinges upon two desirable properties of machine learning models: *generalisation* to unseen domains and *disentanglement* of separable aspects of knowledge.

Generalisation is usually intended theoretically as the gap between train and test error. Neural networks are deemed to excel in this respect (Bottou and Bousquet, 2008) (see also Section 2.2.3). However, this definition has been impoverished in practice by evaluating generalisation only in settings with identically distributed data across splits (Linzen, 2020, *inter alia*). It does not come as a surprise then, that neural networks faced with a distribution shift become unreliable (Rabanser et al., 2019). In fact, rather than relying on the linguistic structures and causal reasoning of humans, they often capture spurious patterns, much like clever Hans (Niven and Kao, 2019). Since cross-lingual transfer may constitute a rather extreme form of distribution shift, it is a better benchmark to assess generalisation.

While my thesis in general is concerned with this challenge, in Chapter 5 I introduced a second, higher-level notion of generalisation which does not concern individual test examples, but rather the *entire target domain*, which is assumed to be previously unseen.

The same way token-level compositionality is entailed in understanding grammatical yet brand-new sentences, in this case a more abstract 'compositionality' should play a role in reassembling relevant knowledge from observed domains. As argued in Chapter 5, this requires the ability '*to disentangle the factors of variation underlying the observed data*' given that they can be separately controlled (Bengio, 2013). Unlocking this ability is key in building machines more similar to humans (Lake et al., 2017).

Moreover, to my knowledge, this is the first work pinpointing the importance of modularity in learning neural models across different domains. In fact, modularity has been shown to be crucial at the level of single episodes of a task, as it achieves better generalisation and robustness to changes in the environment. The ideas at the core of architectures implementing this principle, such as Recurrent Independent Mechanisms (RIMs; Goyal et al., 2019), are: i) sparse interaction among independent modules; ii) soft or hard competition of the modules to become active at every time step to attend to a portion of the sensory input. Goyal et al. (2020) further expanded this framework by enforcing a separation between objects (the states of each module) and schemata (the mechanisms updating them, i.e. the recurrent network parameters).

In a certain way, these two level of modularity (one of RIMs and the other proposed in this thesis) correspond to two levels of memory (Hill et al., 2020; Yogatama et al., 2021). One level is short term and splits the job of tracking how the sensory input changes across time among several processing units. The other kind of memory is long term, and stores general aspects knowledge, which, bundled together in different combinations, are needed to solve different tasks.

6.4 Future Work

Finally, the discussion in Section 6.3 leaves open ample scope for extending the notions of inductive bias and modular design for neural models of language beyond the content of this thesis. In this section, I elaborate on future work along these lines: in addition to offering some detailed ideas for research in natural language processing, I also briefly touch upon other newly opened possibilities for typological linguistics.

6.4.1 A Prior from Emergent Communication

Given the necessity of constructing a prior over neural networks that achieve sample-efficient learning, the idea of cross-lingual knowledge transfer can be pushed even further. Often, not even raw data for a target language are available for unsupervised pre-

training. In this setting, one can exploit artificial languages *emerging* from a referential game on raw images (Kazemzadeh et al., 2014; Lazaridou et al., 2017) as a source for transfer. In particular, artificial agents can be encouraged to cooperate in identifying images among distractors by communicating over vocabularies whose meanings are unknown. The key intuition is that, whereas lexicalisation is mostly arbitrary (Saussure, 1916), communication grounded in a real-world environment (portrayed by images) does constrain what languages are likely or possible (Croft, 2000; Haspelmath, 1999). Hence, I hypothesise that the parameters of a recurrent model that have been optimised for communication over raw images are a favourable starting point to initialise an encoder-decoder model for downstream applications involving genuine natural languages, such as few-shot neural machine translation. Some early results in this direction have already been showcased by Li et al. (2020).

In the past, emergent communication has mostly attracted theoretical interest as a tool to shed light on cooperative behaviours, the compositional properties of emergent communication protocols (Cao et al., 2018; Havrylov and Titov, 2017; Kajić et al., 2020; Lazaridou et al., 2017; Li and Bowling, 2019; Rodríguez Luna et al., 2020), and natural language evolution (Graesser et al., 2019; Kottur et al., 2017). To my knowledge, this would be the first preliminary study on deploying artificial languages from emergent communication in natural language applications. Not only does this hold promise to benefit downstream tasks in resource-learn scenarios in the long run, but also offers an extrinsic evaluation of the properties of languages produced through emergent communication. In particular, one could study the impact of the rate of communication success and maximum sequence length during pre-training on downstream performance.

6.4.2 Parameter Factorisation across Modalities

The success of parameter space factorisation for zero-shot transfer across languages and tasks holds promise to improve generalisation capabilities in other aspects of linguistic knowledge such as multiple modalities (e.g. text, vision, and speech). In fact, state-of-the-art multi-modal neural networks are based on architectures similar to the one outlined in Chapter 4: a Transformer-based encoder creates contextualised representations of text and images, and a task-specific head relies on these representations for classification or regression. For instance ViLBERT (Lu et al., 2019) reserves a separate encoding stream for each modality, and fuses them together at a higher level through a co-attention mechanism, that allows textual tokens to attend to image segments, and vice versa. For instance, given a textual input X_L, a visual input X_V, and H attention heads, in higher layers the Transformer attention mechanism is substituted with:

$$\bigoplus_{i=1}^{H} \left[\text{softmax}\left(\frac{\mathbf{Q}_V \mathbf{x}_V \, \mathbf{x}_L^\top \mathbf{K}_L^\top}{\sqrt{d}} \right) \mathbf{V}_L \mathbf{x}_L \right]_i \mathbf{O}_V \tag{6.1}$$

where $\oplus$ stands for concatenation, and the parameters are specific for queries $\mathbf{Q}$, keys $\mathbf{K}$, values $\mathbf{V}$, and output $\mathbf{O}$ (Vaswani et al., 2017). The other architecture components, included inter-layer FFNs, highway layers, and layer normalisation remain identical.

However, extending this naive approach beyond monolingual English models may lead to a series of problems: 1) The proliferation of parameters without information sharing. If each language is granted a private encoding stream, the number of resulting parameters would multiply accordingly without taking advantage of cross-modal commonalities. 2) Lack of generalisation. After observing two matched inputs of the domains English–vision and English–Tamil, an ideal model should be able to encode Tamil–vision properly, too; however, this is impossible under the existing approaches. Instead, how to share information across modalities and languages, and let the model generalise on unseen combinations?

A possible solution proposed in this thesis is parameter factorisation. In addition to simply conditioning the generation of neural parameters for the classifier head, the concurrence of multiple modalities requires to handle the encoder parameters similarly. In particular, the weights in Equation (6.1) could be graphically dependent on variables representing a specific language l, task t, and modality m, such that $\boldsymbol{\vartheta}_{\text{BERT}} \sim p(\cdot \mid \boldsymbol{l}, \boldsymbol{t}, \boldsymbol{m})$. Following the method proposed in Chapter 5, such probability could be inferred through a hyper-network mapping from variable samples to weight values. However, in this case the cardinality $|\boldsymbol{\vartheta}_{\text{BERT}}|$ is massive, which makes such mapping memory-intensive. To mitigate this problem, rather than generating all encoder parameters, factorisation could be limited to Adapter layers (Houlsby et al., 2019).[3]

6.4.3 Gradient Typology

Finally, the perspective I advocated for, where machine learning is inspired by cross-lingual variation and language acquisition, may help avert some fundamental limitations of the current theoretical approaches to linguistic typology. In fact, typological database documentation is incomplete, approximate, and discrete. As a consequence, it does not fit well with the gradient and contextual models of machine learning.

However, typological databases are originally created from raw linguistic data. Hence, a solution could involve learning typology from such data automatically (i.e. from scratch).

[3] Adapter layers are modules inserted between encoder layers, which remain fixed, after pre-training. They are the only parameters trained during fine-tuning.

This would capture the variation within languages at the level of individual examples, and naturally encode typological information into continuous representations. These goals have already been partly achieved by methods involving language vectors, heuristics derived from morphosyntactic annotation, or distributional information from multi-parallel texts (Ponti et al., 2019a). The main future challenge is the integration of bottom-up typological information into machine learning models, as opposed to sourcing typological features from databases.

Another alternative development of the usage of typological information for natural language processing could take inspiration from the family of methods known as retrofitting or semantic specialisation. These methods inject external lexical knowledge into distributed representations of words (Kamath et al., 2019; Majewska et al., 2020; Mrkšić et al., 2017; Ponti et al., 2018b, 2019c, *inter alia*). Since typological knowledge pertains each language in its entirety, or abstract formal strategies, it is not suitable to enrich word-level embeddings. However, in Chapter 5 I showed how to create *language embeddings*. Hence, typological information could be injected at this level instead to refine language representations learned end-to-end from textual or labelled data.

The possibilities revealed by the experiments presented in the current thesis extend beyond the research ideas detailed in the previous sections. A tighter integration of the processes of learning of humans and machines touches upon several elements that are left for future research. For instance, why are extremely early phases of learning pivotal (Frankle et al., 2020)? Given that stimuli are limited, how important is it to mimic 'parentese', the language mothers and fathers use to address their children, for instance through curriculum learning (Elman, 1993)? Can we shed light on the cognitive substratum of the cross-lingual tendencies in meaning-to-form mapping combining natural language processing and brain imaging techniques (Toneva and Wehbe, 2019)? Hopefully, the insights presented in this thesis will contribute to finding an answer to these open questions.

Bibliography

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of ACL–IJCNLP*, pages 268–272.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the ACL*, 4:301–312.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the ACL*, 4:431–444.

Pascal Amsili and Olga Seminck. 2017. A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29.

Lloyd B. Anderson. 1986. Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In Wallace L. Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*, pages 273–312. Ablex.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the ACL*, 7:597–610.

Adriano Azevedo-Filho and Ross D. Shachter. 1994. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Proceedings of UAI*, pages 28–36.

Dik Bakker. 2010. Language sampling. In J. J. Song, editor, *The Oxford handbook of linguistic typology*, pages 100–127. Oxford University Press.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP*, pages 127–135.

Peter L Bartlett. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536.

Thomas Bayes. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, LII(53):370–418.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 3(6):1–26.

Yoshua Bengio. 2013. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*, pages 1–37.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of ICLR*.

Balthasar Bickel. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

Balthasar Bickel. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine and Heiko Narrog, editors, *Oxford handbook of linguistic analysis*, pages 901–923. Oxford University Press Oxford.

Derek Bickerton. 2015. *Roots of language*. Language Science Press.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020a. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of AAAI*.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of NAACL-HLT*, pages 907–916.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, 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.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. In *Proceedings of ICML*, pages 1613–1622.

Winfried Boeder. 2000. Evidentiality in Georgian. In Lars Johanson and Bo Utas, editors, *Evidentials: Turkic, Iranian and Neighbouring Languages*, volume 24 of *Empirical Approaches to Language Typology*, pages 275–328. Mouton de Gruyter.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Léon Bottou and Olivier Bousquet. 2008. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168.

Melissa Bowerman. 1973. *Early syntactic development: A cross-linguistic study with special reference to Finnish*, volume 11. Cambridge University Press.

Melissa Bowerman. 2011. Linguistic typology and first language acquisition. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

Cecil H. Brown, Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(4):285–308.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z. Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *Proceedings of ICLR*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Noam Chomsky. 1959. A review of BF Skinner's verbal behavior. *Language*, 35(1):26–58.

Noam Chomsky. 1980. On cognitive structures and their development: A reply to Piaget. In Massimo Piattelli-Palmarini, editor, *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Eve V Clark. 1976. Universal categories: On the semantics of classifiers and children's early word meanings. *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday*, 3:449–462.

Eve V Clark. 2001. Emergent categories in first language acquisition. In M. Bowerman and S. C. Levinson, editors, *Language acquisition and conceptual development*, pages 379–405. Cambridge University Press.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDiQA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the ACL*.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the world's languages. http://sswl.railsplayground.net/.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago Press.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Technical report, Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Greville G Corbett. 2010. Implicational hierarchies. In J. J. Song, editor, *The Oxford handbook of linguistic typology*, pages 190–205. Oxford University Press.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of IJNLP*, pages 91–96, Taipei, Taiwan.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of EMNLP*, pages 748–759.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of NAACL-HLT*, pages 536–541.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience.

S Cristofaro and P Ramat. 1999. *Introduzione alla tipologia linguistica*. Carocci.

William Croft. 1995. Autonomy and functionalist linguistics. *Language*, 71(3):490–532.

William Croft. 2000. *Explaining language change: An evolutionary approach*. Pearson Education.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

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

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of TLT*, pages 63–75.

George Cybenko. 1989. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:183–192.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in Artificial Intelligence. *Communications of the ACM*, 58(9):92–103.

Bruno De Finetti. 1929. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of ACL*, pages 399–408.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Amichay Doitch, Ram Yazdi, Tamir Hazan, and Roi Reichart. 2019. Perturbation based learning for structured NLP tasks with application to dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:643–659.

Matthew S Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(2):257–292.

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

John Duchi. 2007. Derivations for linear algebra and optimization. Technical report, University of California, Berkeley.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of ACL*, pages 845–850.

William H Durham. 1991. *Coevolution: Genes, culture, and human diversity*. Stanford University Press.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of EMNLP–CoNLL*, pages 1–11.

Gintare Karolina Dziugaite and Daniel M. Roy. 2015. Neural network matrix factorization. *arXiv preprint arXiv:1511.06443*.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21.

Nicholas Evans. 2011. Semantic typology. In J. J. Song, editor, *The Oxford Handbook of Linguistic Typology*, pages 504–533. Oxford University Press.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, pages 1126–1135.

JR Firth. 1957. Modes of meaning. In *Papers in Linguistics 1934-1951*. Oxford University Press.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. 2018. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of ICML*, pages 1568–1577.

Jonathan Frankle, David J Schwab, and Ari S Morcos. 2020. The early phase of neural network training. In *International Conference on Learning Representations*.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Yarin Gal and Zoubin Ghahramani. 2016a. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of ICML*, pages 1050–1059.

Yarin Gal and Zoubin Ghahramani. 2016b. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of NeurIPS*, pages 1019–1027.

Andrew Gelman, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of COLING*, pages 385–393.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018a. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association of Computational Linguistics*, 6:451–465.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018b. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP*, pages 316–327.

Gary Martin Gilligan. 1989. *A cross-linguistic approach to the pro-drop parameter*. Ph.D. thesis, University of Southern California.

Pam Frost Gorder. 2006. Neural networks show new promise for machine vision. *Computing in science & engineering*, 8(6):4–8.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, volume 46.

Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Sergey Levine, Charles Blundell, Yoshua Bengio, and Michael Mozer. 2020. Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems. *arXiv preprint arXiv:2006.16225*.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. 2019. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*.

Laura Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of EMNLP-IJCNLP*, pages 3698–3708.

Giorgio Graffi. 1980. Universali di Greenberg e grammatica generativa in la nozione di tipo e le sue articolazioni nelle discipline del linguaggio. *Lingua e Stile Bologna*, 15(3):371–387.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2017. Improving neural language models with a continuous cache. In *Proceedings of ICLR*.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Joseph H Greenberg. 1966. *Universals of language*. MIT Press.

Ferdinand de Haan. 2013. Coding of evidentiality. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Michael Alexander Kirkwood Halliday. 1975. *Learning how to mean*. Hodder Arnold.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog 4.2.1. Max Planck Institute for the Science of Human History.

Zellig S Harris. 1951. *Methods in structural linguistics*. University of Chicago Press.

Martin Haspelmath. 1993. *A Grammar of Lezgian*, volume 9 of *Mouton Grammar Library*. Mouton de Gruyter.

Martin Haspelmath. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2):180–205.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Micheal Tomasello, editor, *The new psychology of language*, volume 2, pages 1–30. Psychology Press.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic typology*, 11(1):119–132.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Proceedings of NeurIPS*, pages 2149–2159.

Marvin I Herzog, William Labov, and Uriel Weinreich. 1968. Empirical foundations for a theory of language change. In WP Lehmann and Y. Malkiel, editors, *Directions for Historical Linguistics*, pages 95–195. University of Texas Press.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow. *arXiv preprint arXiv:2009.01719*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Paul J. Hopper. 1979. Aspect and foregrounding in discourse. In *Discourse and Syntax*, pages 211–241. Brill.

Kurt Hornik. 1991. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML*.

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*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Ivana Kajić, Eser Aygün, and Doina Precup. 2020. Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint arXiv:2004.01097*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP*, pages 1700–1709.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 72–83.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of ACL*, pages 1993–2003.

Paul Kay and Chad K McDaniel. 1978. The linguistic significance of the meanings of basic color terms. *Language*, 54(3):610–646.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of EMNLP*, pages 787–798.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of NeurIPS*, pages 5574–5584.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proceedings of ICLR*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of ICML*, pages 595–603.

Max Kochurov, Timur Garipov, Dmitry Podoprikhin, Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2018. Bayesian incremental learning for deep neural networks. In *Proceedings of ICLR (Workshop Papers)*.

András Kornai. 2013. Digital language death. *PloS One*, 8(10):e77056.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of EMNLP*, pages 2962–2967.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of NeurIPS*, pages 6402–6413.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Pierre Simon Laplace. 1820. *Théorie analytique des probabilités*. Courcier.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of ICLR*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

M Paul Lewis, Gary F Simons, and Charles D Fennig. 2016. *Ethnologue: Languages of the world*, 19th edition. SIL international.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. In *Proceedings of NeurIPS*, pages 15825–15835.

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2020. Emergent communication pretraining for few-shot machine translation. In *Proceedings of COLING*, pages 4716–4731.

David W Lightfoot. 1979. Principles of diachronic syntax. *Cambridge University Press*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Björn Lindblom. 1986. Phonetic universals in vowel systems. In Jaeger Ohala, editor, *Experimental Phonology*, pages 13–44. Academic Press.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of ACL*, pages 5210–5217.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL*, pages 8–14.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019. DARTS: Differentiable architecture search. In *Proceedings of ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in NeurIPS*, pages 13–23.

Jelena Luketina, Mathias Berglund, Klaus Greff, and Tapani Raiko. 2016. Scalable gradient-based tuning of continuous regularization hyperparameters. In *Proceedings of ICML*, pages 2952–2960.

Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. 2008. SoRec: Social recommendation using probabilistic matrix factorization. In *Proceedings of CIKM*, pages 931–940.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

David JC MacKay. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. 2015. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of ICML*, pages 2113–2122.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of ICLR*.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo M Ponti, and Anna Korhonen. 2020. Verb knowledge injection for multilingual event processing. *arXiv preprint arXiv:2012.15421*.

Asifa Majid, Melissa Bowerman, Miriam van Staden, and James S Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–152.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of EMNLP*, pages 2529–2535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of EMNLP*, pages 1192–1202.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

R Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*.

Garland McNew, Curdin Derungs, and Steven Moran. 2018. Towards faithfully visualizing global linguistic diversity. In *Proceedings LREC*.

Gabriela Melo, Vinicius Imaizumi, and Fábio Cozman. 2020. Esquemas de Winograd em português. In *Proceedings of the 16th Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2020)*, pages 787–798.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of ACL*, pages 4975–4989.

Matti Miestamo. 2004. *Clausal negation: A typological study.* Ph.D. thesis, University of Helsinki.

Tom M. Mitchell. 1980. The need for biases in learning generalizations. Technical report, Rutgers Computer Science. CBM-TR-117.

Andriy Mnih and Ruslan Salakhutdinov. 2008. Probabilistic matrix factorization. In *Proceedings of NeurIPS*, pages 1257–1264.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*. MIT Press.

Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Andrea Moro and Noam Chomsky. 2015. *The boundaries of Babel: The brain and the enigma of impossible languages*. MIT Press.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association of Computational Linguistics*, 5(1):309–324.

Tsendsuren Munkhdalai and Adam Trischler. 2018. Metalearning with Hebbian fast weights. *arXiv preprint arXiv:1807.05076*.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT Press.

Radford M Neal. 1996. *Bayesian learning for neural networks*, volume 118. Springer–Verlag.

Radford M. Neal. 2011. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–163. Routledge.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. 2019. The role of over-parametrization in generalization of neural networks. In *Proceedings of ICLR*.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. 2018. Variational continual learning. In *Proceedings of ICLR*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of ACL*, pages 4658–4664.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, et al. 2019. Universal Dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Peter Orbanz. 2012. Lecture notes on Bayesian nonparametrics. *Journal of Mathematical Psychology*, 56:1–12.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the EACL*, pages 644–649.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL*, volume 1, pages 1946–1958.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of ICML*, pages 1310–1318.

Fabian Pedregosa. 2016. Hyperparameter optimization with approximate gradient. In *Proceedings of ICML*, pages 737–746.

Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the NAACL–HLT*, pages 2227–2237.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.

Gabriel Peyré. 2020. Course notes on optimization for machine learning. Technical report, CNRS, DMA, École normale supérieure.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of EMNLP*, pages 425–435.

Edoardo M Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*. To appear.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP*, pages 2362–2376.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018a. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of ACL*, pages 1531–1542.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. Towards zero-shot language modeling. In *Proceedings of EMNLP*, pages 2900–2910.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018b. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019c. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP-IJCNLP*, pages 2206–2217.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. In *Proceedings of NeurIPS*, pages 1394–1406.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of ACL*, pages 151–164.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of ACL*, pages 463–473.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of NAACL-HLT*.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? The case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *Proceedings of ICLR*.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime G. Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of AAAI*, pages 6924–6931.

Jan Rijkhoff, Dik Bakker, Kees Hengeveld, and Peter Kahrel. 1993. A method of language sampling. *Studies in Language*, 17(1):169–203.

Hippolyt Ritter, Aleksandar Botev, and David Barber. 2018. Online structured Laplace approximations for overcoming catastrophic forgetting. In *Proceedings of NeurIPS*, pages 3738–3748.

Diana Rodríguez Luna, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni. 2020. Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of EMNLP*, pages 4428–4437.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 2011 AAAI Spring Symposium Series*.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. KLcpos3 - A language similarity measure for delexicalized parser transfer. In *Proceedings of ACL*, pages 243–249.

Malcolm Ross. 1997. Social networks and kinds of speech community event. In Roger M. Blench and Matthew Spriggs, editors, *Archaeology and Language, I*, pages 209–261. Routledge.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019a. Transfer learning in natural language processing. In *Proceedings of NAACL-HLT: Tutorials*, pages 15–18.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019b. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial Winograd Schema Challenge at scale. In *Proceedings of AAAI*, pages 8732–8734.

Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of ICML*, pages 880–887.

Reginaldo J Santos. 1996. Equivalence of regularization and truncated iteration for general ill-posed problems. *Linear algebra and its applications*, 236:25–33.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of EMNLP-IJCNLP*, pages 4463–4473.

Edward Sapir. 2014 [1921]. *Language*. Cambridge University Press.

Ferdinand de Saussure. 1916. *Cours de linguistique générale, ed.* Payot. Edited by C. Bally and A. Sechehaye.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the EMNLP-IJCNLP*, pages 5835–5841.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *Proceedings of ICML*, pages 4528–4537.

Hanhuai Shan and Arindam Banerjee. 2010. Generalized probabilistic matrix factorizations for collaborative filtering. In *Proceedings of ICDM*, pages 1025–1030.

Ehsan Shareghi, Yingzhen Li, Yi Zhu, Roi Reichart, and Anna Korhonen. 2019. Bayesian learning for neural dependency parsing. In *Proceedings of NAACL-HLT*, pages 3509–3519.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2015. 日本語 Winograd Schema Challenge の構築と分析. 言語処理学会第 *21* 回年次大会論文集, pages 493–496.

Yoav Shoham. 1990. Nonmonotonic reasoning and causation. *Cognitive Science*, 14(2):213–252.

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language*, 31(4):507–524.

Dan I. Slobin. 1973. Cognitive prerequisites for the development of grammar. In Charles Ferguson and Dan I. Slobin, editors, *Studies of child language development*, pages 175–208. Holt, Rinehart, & Winston.

Dan I Slobin. 1985. Crosslinguistic evidence for the language-making capacity. *The crosslinguistic study of language acquisition*, 2:1157–249.

Paul Smolensky and Alan Prince. 1993. *Optimality Theory: Constraint interaction in generative grammar*. Wiley Online Library.

Benjamin Snyder and Regina Barzilay. 2010. Climbing the tower of Babel: Unsupervised multilingual learning. In *Proceedings of ICML*, pages 29–36.

Richard Sproat. 2016. Language typology in speech and language technology. *Linguistic Typology*, 20(3):635–644.

S. N. Sridhar. 1990. *Kannada: Descriptive Grammar*. Croom Helm Descriptive Grammars. Routledge.

Jake Stolee and Neill Patterson. 2019. Matrix factorization with neural networks and stochastic variational inference. Technical report, University of Toronto.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL-HLT*, pages 477–487.

Corentin Tallec and Yann Ollivier. 2017. Unbiasing truncated backpropagation through time. *arXiv preprint arXiv:1705.08209*.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of ACL*, pages 4911–4921.

Leonard Talmy. 1983. How language structures space. In H. L. Pick Jr., editor, *Spatial orientation*, pages 225–282. Springer.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with Universal Dependencies and predicted POS labels. *Proceedings of Depling*, pages 340–349.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14954–14964.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of NAACL-HLT*, pages 1357–1366.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*, volume 1, pages 1661–1670.

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 *Proceedings of NeurIPS*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of NIPS*, pages 3630–3638.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of neural networks using DropConnect. In *Proceedings of ICML*, pages 1058–1066.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Larry Wasserman. 2013. *All of statistics: a concise course in statistical inference*. Springer.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Andrew Gordon Wilson. 2019. The case for Bayesian deep learning. *NYU Courant Technical Report*.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. 2019. SNAS: Stochastic neural architecture search. In *Proceedings of ICLR*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, pages 1–8.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *arXiv preprint arXiv:2102.02557*.

Sandy L Zabell. 2005. *Symmetry and its discontents: essays on the history of inductive probability*. Cambridge University Press.

Anthony M Zador. 2019. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP*, pages 35–42.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of ICLR*.

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *Proceedings of EMNLP*, pages 1368–1378.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of EMNLP*, pages 238–249.

Appendix A

Background

A.1 Activation Functions and Derivatives

	$\phi(x)$	$\frac{d\phi(x)}{dx}$
ReLU	$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$	$\begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$
Tanh	$\frac{e^x - e^{-x}}{e^x + e^{-x}}$	$1 - \phi(x)^2$
Sigmoid	$\frac{1}{1 + e^{-x}}$	$\phi(x)(1 - \phi(x))$
	$\phi(\mathbf{x}_i)$	$\frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_j}$
Softmax	$\frac{e^{x_1}}{\sum_j e^{x_j}}$	$\phi(\mathbf{x}_i)(\delta_{i,j} - \phi(\mathbf{x}_j))$

Table A.1 Common element-wise (top) and array-wise (bottom) non-linear activation functions (left) and their total or partial derivatives (right). δ_{ij} is the Kronecker delta: it equals 1 if $i = j$ and 0 if $i \neq j$.

Appendix B

A Prior over Weights for Language Modelling

B.1 List of iso 639-3 codes and language names

In Table B.1, the ISO 639-3 codes for each language are associated with the corresponding language name. In addition to this information, Table B.1 provides the total count of characters for the three data splits, and the type-to-token ratio.

B.2 Typological Features

The 245 binarized typological features from Littell et al. (2017) that define the general properties of each language are plotted as a heat map in Figure B.1. Features are related to syntax if their name starts with S, to phonology if it starts with P, and to phonemic inventories if it starts with INV. Note how some values are so rare that they belong exclusively to a single language in the sample, e.g. the vowel /ɘ/ for Thai.

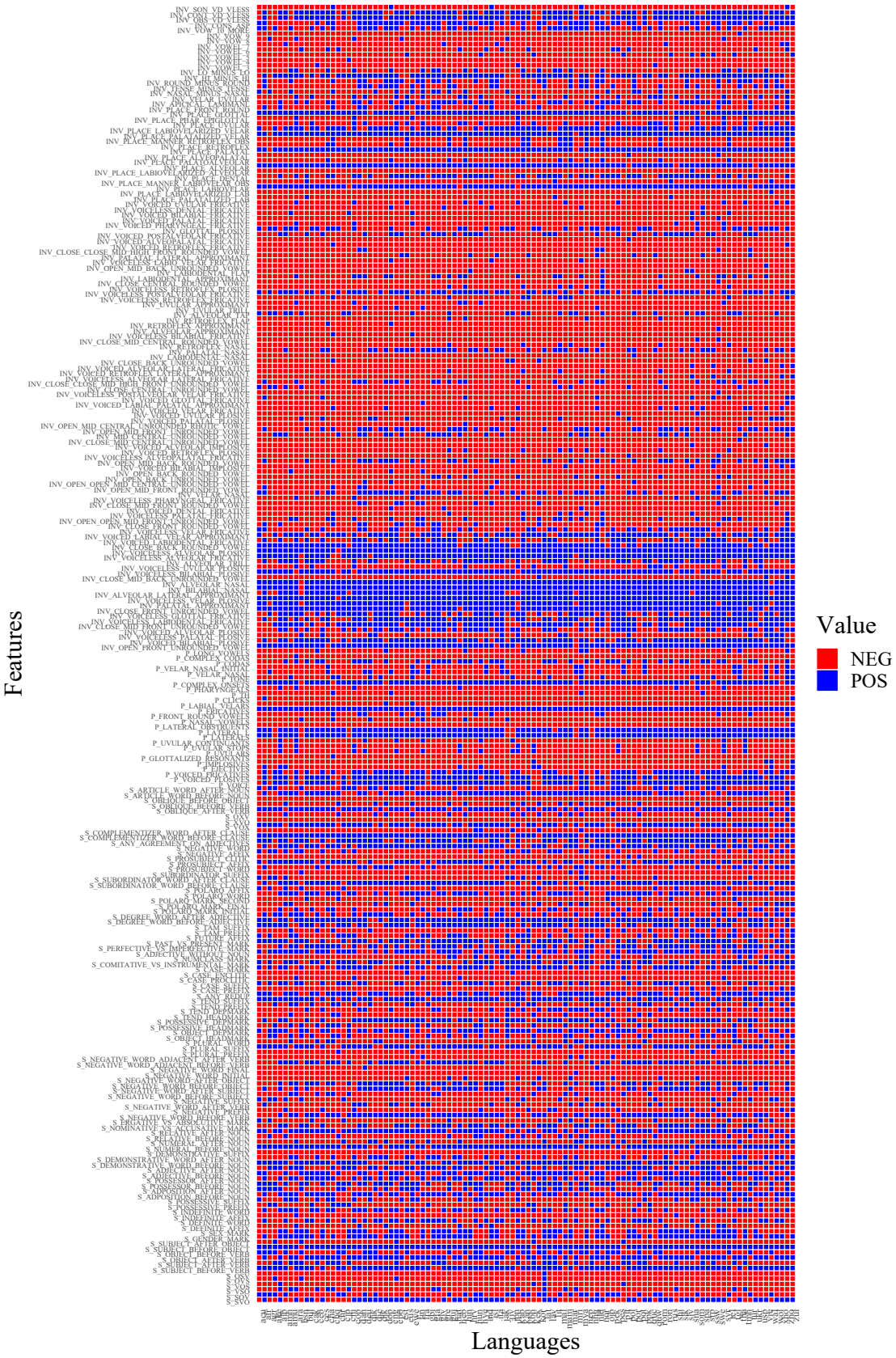


Figure B.1 Binary values of the typological features from Littell et al. (2017) (y-axis) for each language (x-axis).

ISO	Name	Train char	Dev char	Eval char	Typ / Tok
acu	Achuar	1149777	136773	119849	5.902^{-05}
afr	Afrikaans	3229549	413064	437532	1.985^{-05}
agr	Aguaruna	991098	118726	103237	6.348^{-05}
ake	Akawaio	960849	113905	111793	5.141^{-05}
alb	Albanian	3152312	402399	427612	1.808^{-05}
amu	Amuzgo	1156128	142241	132194	5.662^{-05}
bsn	Barasana	1397953	171482	162042	4.736^{-05}
cak	Cakchiquel	1404839	169031	161609	4.033^{-05}
ceb	Cebuano	3985326	509809	536615	1.471^{-05}
ces	Czech	2756308	349505	371027	2.675^{-05}
cha	Chamorro	641087	66469	67935	1.032^{-04}
chq	Chinantec	1548993	174921	164087	4.502^{-05}
cjp	Cabecar	856441	100035	97246	8.256^{-05}
cni	Campa	1149737	133104	120600	3.990^{-05}
dan	Danish	2774922	364278	385352	2.298^{-05}
deu	German	3195266	391700	417235	2.023^{-05}
dik	Dinka	716411	84429	81572	6.800^{-05}
dje	Zarma	3126629	372921	405747	1.767^{-05}
djk	Aukan	1083303	129770	124682	5.083^{-05}
dop	Lukpa	864347	96068	95094	6.442^{-05}
eng	English	3238389	418697	450324	1.509^{-05}
epo	Esperanto	3029361	391874	409980	1.514^{-05}
est	Estonian	778681	177680	5329	8.007^{-05}
eus	Basque	801072	94904	93712	7.073^{-05}
ewe	Ewe	870990	97081	94738	9.315^{-05}
fin	Finnish	3195802	402386	426808	1.789^{-05}
fra	French	3246315	404464	435820	2.129^{-05}
gbi	Galela	1347199	144215	129606	4.442^{-05}
gla	Gaelic	68110	6802	7167	5.848^{-04}
glv	Manx	392897	58690	48239	1.360^{-04}
hat	Creole	3332162	400606	440487	1.989^{-05}
hrv	Croatian	2594494	340781	359694	2.549^{-05}
hun	Hungarian	3020721	376738	408697	2.391^{-05}
ind	Indonesian	3528757	405822	454277	1.823^{-05}
isl	Icelandic	2968652	376562	389091	2.517^{-05}
ita	Italian	2979890	388587	409629	2.091^{-05}
jak	Jakalteko	1116611	131793	122853	5.615^{-05}
jiv	Shuar	888886	98309	97483	5.624^{-05}
kab	Kabyle	798503	91677	87964	7.770^{-05}
kbh	Camsa	1373946	144068	140039	5.307^{-05}
kek	Q'eqchi'	4375494	525831	517455	1.735^{-05}
lat	Latin	2700731	325553	342365	1.514^{-05}
lav	Latvian	644923	77572	78263	1.161^{-04}
lit	Lithuanian	2531703	313391	309237	2.536^{-05}
mam	Mam	1053107	119781	112869	6.222^{-05}
mri	Maori	3504361	437499	456364	1.501^{-05}
nhg	Nahuatl	1126416	135355	126213	5.548^{-05}
nld	Dutch	3224058	392079	416432	2.033^{-05}
nor	Norwegian	2941245	374161	392574	2.508^{-05}
pck	Paite	3174091	404462	401042	1.784^{-05}
plt	Malagasy	3744462	468678	477671	1.705^{-05}
pol	Polish	2963005	374471	398263	2.088^{-05}
por	Portuguese	3010541	380551	404559	2.450^{-05}
pot	Potawatomi	212243	25336	24020	1.911^{-04}
ppk	Uma	1050858	115947	110127	5.090^{-05}
quc	K'iche'	1153252	131623	127281	5.382^{-05}
quw	Quichua	792834	93791	90930	8.593^{-05}
rom	Romani	818094	91328	89036	7.912^{-05}
ron	Romanian	3107966	394280	419241	1.760^{-05}
shi	Tachelhit	738833	82470	79965	5.659^{-05}
slk	Slovak	2821379	361684	385812	2.662^{-05}
slv	Slovene	2883854	369397	381124	2.366^{-05}
sna	Shona	3013970	384548	407708	2.154^{-05}
som	Somali	3750398	468849	498051	1.314^{-05}
spa	Spanish	3082800	388079	410641	2.190^{-05}
srp	Serbian	2503088	319878	341496	2.402^{-05}
ssw	Swahili	763781	84855	82704	6.979^{-05}
swe	Swedish	3206128	403712	427604	1.932^{-05}
tgl	Tagalog	3848347	477839	505730	1.283^{-05}
tmh	Tuareg	270666	30146	31636	2.436^{-04}
tur	Turkish	2719803	318179	345666	2.601^{-05}
usp	Uspanteco	1134539	131631	125891	5.747^{-05}
vie	Vietnamese	3194226	379697	417496	4.635^{-05}
wal	Wolaytta	837506	98141	96951	6.004^{-05}
wol	Wolof	683480	80261	77575	8.201^{-05}
xho	Xhosa	3005476	377342	399338	1.692^{-05}
zul	Zulu	690644	80944	77975	8.946^{-05}

Table B.1 Stats for all languages part of the corpus: for each ISO 639-3 code, I list (from left to right) the language name, the character count (in the train, development, and evaluation sets), and the type-to-token ratio.

Appendix C

A Prior over Architectures for Language Understanding

C.1 Detailed Translation Guidelines

Translation of the English COPA validation and test set instances into each of the 11 languages was carried out by a single translator per language, meeting the following eligibility criteria: (i) a native speaker of the target language, (ii) fluent in English, (iii) with minimum undergraduate education level. Each translator was presented with translation guidelines and a spreadsheet accessible online, containing one English premise-prompt-hypothesis triple per line, followed by an empty line where target translations were entered. The task consisted in (a) identifying the correct alternative for the English premise and (b) translating the premise and both alternative hypotheses into the target language, preserving the causal relations present in the original. Each translator worked independently (using any external resources, such as English-target language dictionaries, if needed) and completed the task in its entirety, producing 100 validation and 500 test instance translations, and a label for each. To ensure the output preserves the lexical, temporal, and causal relations present in the original triples, the guidelines instructed to:

1. maintain the original correspondence relations between lexical items, i.e., if the same English word appeared both in the premise and the alternatives (Premise: *The friends decided to share the hamburger.*; A1: *They cut the hamburger in half.*; A2: *They ordered fries with the hamburger.*), it was translated into the same target-language equivalent in all three translated sentences;

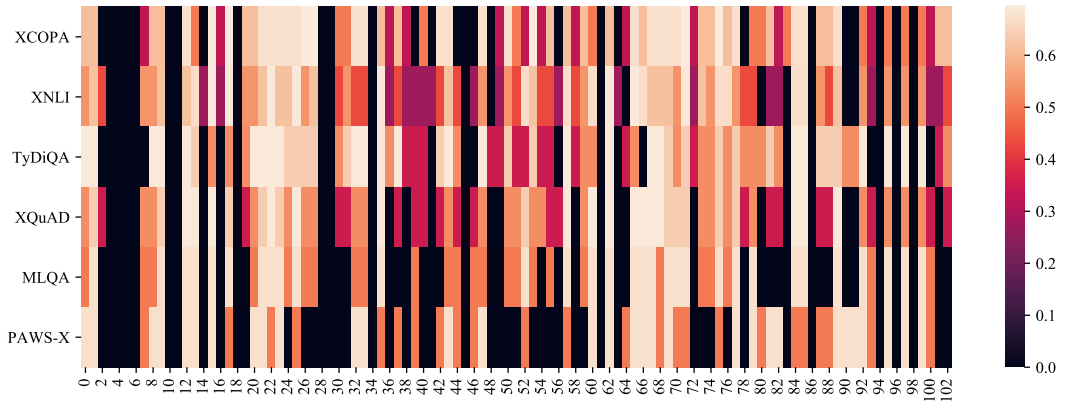


Figure C.1 Heatmap of the entropy of the distributions of WALS features (x axis) in language samples from famous cross-lingual datasets outlined in Section 5.6 (y axis).

2. ensure that the original chronology and temporal extension of events is preserved through appropriate choice of verbal tense and aspect in the target language, e.g., maintaining the distinction between perfective and imperfective aspect (Premise: *My eyes became red and puffy.* [PERF], A1: *I was sobbing.* [IMPERF], A2: *I was laughing.* [IMPERF];

3. in case of English words with no exact translations in the target language or referring to concepts absent from the target language culture (e.g., *peach*), the following solutions were to be adopted, in order of preference: (1) using a common loanword from another language, provided it is understood by the general population of target-language speakers; (2) using a periphrasis to describe the same concept (e.g., *a juicy fruit*); (3) substituting the original concept with a similar one that is more familiar to the target language speaker community (e.g., *santol*), provided that it can play a similar role in the causal relations captured by the original premise-prompt-hypothesis triple;

The translators were encouraged to split the workload into multiple sessions with breaks in between. On average, the task took 20 hours of work to complete. Additionally, translators were encouraged to provide feedback, commenting on translation challenges and chosen solutions.

Setup	Model	EN	ET	HT	ID	IT	QU	SW	TA	TH	TR	VI	ZH
CO-ZS	XLM-R	57.6	59.8	49.4	58	56	50.7	57.2	56.6	52.8	56.2	58.5	56.6
	XLM-R-L	53	49.6	55.8	53	52.4	48	54	51.4	51.8	51	56	53
	MBERT	62	50.6	51.4	55	53.8	54.7	53.6	52	53.2	56.8	55.4	59
	USE	63	53.8	49.4	57.6	60	48.3	52.2	53	57.2	55	54.8	60.2
CO-TLV	XLM-R	57.6	57.8	48.6	60.8	54.4	49.5	55.4	55.8	54.2	54.8	57.6	57.2
	XLM-R-L	53	49.4	47.8	51.4	53.6	54.2	50	47.8	53	50.6	58.2	51
	MBERT	62	52	52.6	58.2	55	52.7	53	52	52.4	53.8	52.6	61.8
	USE	63	49.4	49.6	57.6	62	54	50.8	53.6	58.6	56.2	51.4	59.2
SI-ZS	XLM-R	68	59.4	49.2	67.2	63.6	51	57.6	58.8	61.6	60.4	65.8	66
	XLM-R-L	85	70.4	53.4	79.4	72.8	50.2	60.8	71	69.4	71.2	76	78.2
	MBERT	62.2	55.2	51.4	57	57	50.2	51	52.2	51	53.2	59.2	64.4
	USE	62.6	51.6	46.8	60.2	61.8	50.5	52.4	48.8	60.8	54.6	54.8	63
SI+CO-ZS	XLM-R	66.8	58	51.4	65	60.2	51.2	52	58.4	62	56.6	65.6	68.8
	XLM-R-L	84.2	68.8	52.8	79.8	72.4	50.7	59.4	68.2	67.2	71.2	73.8	76.2
	MBERT	63.2	52.2	54	59.4	57.2	48	56	54.6	51.2	57.4	58	65.6
	USE	63.8	51.2	48.4	57.6	61.8	52	51.8	47	58	55.6	51	60.2
SI+CO-TLV	XLM-R	66.8	59.4	50	71	61.6	46	58.8	60	63.2	62.2	67.6	67.4
	XLM-R-L	84.2	71.4	52.8	79.8	72.6	52	59.2	73	72.8	74.4	73.8	78.6
	MBERT	63.2	52.2	51.8	58.2	57.2	53	51	57.2	52.6	54.6	57.8	52.4
	USE	63.8	51.8	47.8	56.6	61.6	52.2	52.4	47	59.8	54.4	52.8	60.6

Table C.1 Detailed per-language XCOPA results. None of the models was exposed to HT and QU in pretraining. USE was exposed in pretraining only to IT, TH, TR, and ZH.

C.2 Grammatical Tense and Aspect in Translation

The scenarios included in COPA refer to events that took place in the past and are formulated in what can be described as a narrative register (one of the sources from which question topics were drawn was a corpus of personal stories published online (Gordon and Swanson, 2009)). This is grammatically rendered exclusively by means of past simple (preterite) or past continuous (imperfect) verb forms. Temporal anteriority of a hypothesis sentence with respect to the premise is not grammatically marked (e.g., with a past perfect verb form) and can only be deduced based on the prompt ("*What was the* CAUSE *of this?*"). The preterite-imperfect contrast used in English to distinguish background states (imperfective) from the main event (perfective) (e.g., *I was expecting company.* IMPERF vs. *I tidied up my house.* PERF) is not universally applicable and different languages employ different discourse grounding strategies (Hopper, 1979), which has interesting implications for the multilingual extension of COPA to XCOPA.

In the languages with grammatical tense different strategies are employed to capture the perfective-imperfective distinction, which is prominent in COPA. For example, in Haitian Creole, the simple past marker *te* is used to indicate a bounded event in the past, while the continuous aspect is signaled with an *ap* marker. Italian additionally distinguishes between two perfective past tenses, expressed by means of a simple and compound past (*vidi* - *ho visto*, 'I saw'). The opposition is between completed actions whose effects are detached from the present and those with persisting effects on the present. Both contrast with the imperfect, which emphasises the event's extension or repetition in time. Given that the opposition is a matter of the speaker's perspective on events rather than based on deixis (remote versus proximate past), the translator opted for the most natural choice given a specific context/situation.

C.3 Hyper-Parameter Search

For MBERT and XLM-R I searched the following hyperparameter grid in both SIQA and COPA training: learning rate $\in \{5 \cdot 10^{-6}, 10^{-5}, 3 \cdot 10^{-5}\}$, dropout rate (applied to the output layer of the transformer and the hidden layer of the feed-forward scoring net) $\in 0, 0.1$, and batch size $\in \{4, 8\}$. For USE, I searched over different values for the learning rate, $\{10^{-3}, 10^{-4}, 10^{-5}\}$. I evaluated the performance on the respective development set every 500 updates for SIQA and every 10 updates for COPA and stopped the training if there was no improvement over 10 consecutive evaluations. In all setups, I optimised the parameters with the Adam algorithm (Kingma and Ba, 2015) ($\epsilon = 10^{-8}$, no weight decay nor warmup) and clipped the norms of gradients in single updates to 1.0.

C.4 Full Results (Per Language)

Table C.1 contains the detailed per language results for all XCOPA languages and all five of the evaluation setups (CO-ZS, CO-TLV, SI-ZS, SI+CO-ZS, SI+CO-TLV).

Name	Lang	Vocab	Params	URL
mBERT	Multiling.	119K	125M	https://huggingface.co/bert-base-multilingual-cased
XLM-R	Multiling.	250K	125M	https://huggingface.co/xlm-roberta-base
XLM-R-L	Multiling.	250K	355M	https://huggingface.co/xlm-roberta-large

Table C.2 Pretrained transformers used in Chapter 4.

C.5 Code and Dependencies

The code is built on top of the HuggingFace Transformers framework: https://github.com/huggingface/transformers. Table C.2 details the LM-pretrained transformer models from this framework which I exploited in this work. For the experiments with USE, I encoded the sequences with the pretrained multilingual (16 languages) encoder available from: https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3. Besides the Transformers library and USE, the code only relies on standard Python's scientific computing libraries (e.g., `numpy`).