Emergent Communication Pretraining for Few-Shot Machine Translation

Yaoyiran Li, Edoardo M. Ponti, Ivan Vulić and Anna Korhonen
Language Technology Lab, TAL, University of Cambridge
{yl711,ep490,iv250,alk23}@cam.ac.uk

Abstract

While state-of-the-art models that rely upon massively multilingual pretrained encoders achieve sample efficiency in downstream applications, they still require abundant amounts of unlabelled text. Nevertheless, most of the world's languages lack such resources. Hence, we investigate a more radical form of unsupervised knowledge transfer in the absence of linguistic data. In particular, for the first time we pretrain neural networks via emergent communication from referential games. Our key assumption is that grounding communication on images—as a crude approximation of real-world environments—inductively biases the model towards learning natural languages. On the one hand, we show that this substantially benefits machine translation in few-shot settings. On the other hand, this also provides an extrinsic evaluation protocol to probe the properties of emergent languages *ex vitro*. Intuitively, the closer they are to natural languages, the higher the gains from pretraining on them should be. For instance, in this work we measure the influence of communication success and maximum sequence length on downstream performances. Finally, we introduce a customised adapter layer and annealing strategies for the regulariser of maximum-a-posteriori inference during fine-tuning. These turn out to be crucial to facilitate knowledge transfer and prevent catastrophic forgetting. Compared to a recurrent baseline, our method yields gains of $59.0\%\sim147.6\%$ in BLEU score with only 500 NMT training instances and $65.1\%\sim196.7\%$ with $1,000$ NMT training instances across four language pairs. These proof-of-concept results reveal the potential of emergent communication pretraining for both natural language processing tasks in resource-poor settings and extrinsic evaluation of artificial languages.

1 Introduction

Zero-shot and few-shot learning are notoriously challenging for neural networks (Bottou and Bousquet, 2008; Vinyals et al., 2016; Ravi and Larochelle, 2017). However, they are a prerequisite for natural language processing in most languages, which suffer from the paucity of annotated data (Ponti et al., 2019a). State-of-the-art models rely on knowledge transfer, whereby an encoder is pretrained via language modeling on texts from multiple languages, and subsequently 'fine-tuned' on labelled examples of resource-rich languages (Wu and Dredze, 2019; Conneau et al., 2020) or few examples in a target resource-poor language (Lauscher et al., 2020). However, even raw texts required for pretraining are scant (Kornai, 2013): for instance, Wikipedia dumps cover 278 languages out of the 7,097 spoken world-wide (Eberhard et al., 2020).

For this reason, we push the idea of cross-lingual knowledge transfer even further, exploring and profiling a setting where not even raw natural language data for a target language are available for unsupervised pretraining. In their stead, we exploit artificial languages *emerging* from a referential game on raw images (Kazemzadeh et al., 2014; Lazaridou et al., 2017). In particular, we encourage agents 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 does constrain what languages are likely or

Proceedings of the 28th International Conference on Computational Linguistics, pages 4716–4731
Barcelona, Spain (Online), December 8-13, 2020

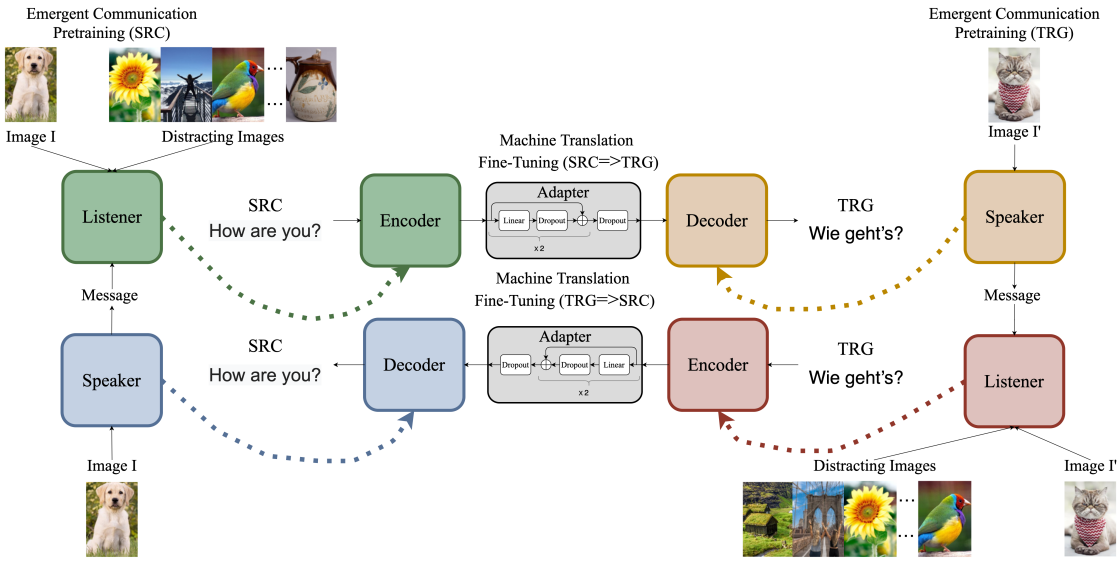

Figure 1: An overview of the model architecture. Dashed lines denote parameter transfer from the EC pretraining task to the MT fine-tuning task. We stress that during EC pretraining, we do not leverage any image-caption pairs; instead, only unlabelled images are used. During MT fine-tuning, standard seq2seq NMT models are trained on SRC and TRG sentence pairs without any visual information available.

possible (Haspelmath, 1999; Croft, 2000). Hence, we hypothesise that communication over raw images offers a favourable inductive bias for natural language tasks.

In particular, we experiment with initialising an encoder-decoder model for few-shot neural machine translation with parameters pretrained on emergent communication. 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 (Lazaridou et al., 2017; Havrylov and Titov, 2017; Cao et al., 2018; Li and Bowling, 2019; Kajić et al., 2020), and natural language evolution (Kottur et al., 2017; Graesser et al., 2019). To our knowledge, this is the first preliminary study on deploying artificial languages from emergent communication in natural language applications.

Conversely, our method also constitutes an extrinsic evaluation protocol to probe the properties of different emergent languages. The underlying assumption is that they should facilitate downstream tasks only to the extent that they share common characteristics with natural languages. In particular, we run in-depth analyses on the impact that the rate of communication success and maximum sequence length have on NMT performance.

For the sake of fully leveraging the pretrained parameters and ameliorating overfitting, we also explore several new strategies to perform knowledge transfer. In particular, we customise the adapter layer (Houlsby et al., 2019) and propose annealing strategies for the regularisation term of MAP inference during fine-tuning. We run experiments in NMT between English and four languages (German, Czech, Romanian, and French) in both directions. By virtue of emergent communication pretraining and the proposed transfer strategies, we report gains in BLEU scores when simulating few-shot MT setups for the four target languages: $59.0\%\sim147.6\%$ over a standard encoder-decoder baseline when 500 training instances are available, and $65.1\%\sim196.7\%$ when $1,000$ training instances are available. Our code is available online at `https://github.com/cambridgeltl/ECNMT`.

2 Related Work

Our work lies at the intersection of several prominent research areas such as pretraining for transfer learning, emergent communication, few-shot machine translation, and inductive biases for language. To all of these we cannot do full justice given space constraints.

Pretraining for Transfer Learning. Unsupervised pretraining on large collections of unlabelled text yields general-purpose contextualized word representations (Peters et al., 2018; Howard and Ruder, 2018)

that are beneficial across a range of downstream NLP tasks. The current dominant paradigm is training a Transformer-based deep model (Vaswani et al., 2017) relying on masked language modeling or a similar objective, as proposed in the omnipresent BERT model (Devlin et al., 2019) and its extensions (Liu et al., 2019; Conneau and Lample, 2019; Song et al., 2019; Joshi et al., 2020), and then fine-tuning the model further on a downstream task (Wang et al., 2019).

Often this approach exploits large textual data and deep models spanning even billions of parameters (Conneau et al., 2020; Raffel et al., 2019; Brown et al., 2020). In this work, we refrain from chasing task leaderboards (Linzen, 2020) and posit a fundamental question about language learning instead.

Emergent Communication. The functional aspect of language (Clark, 1996) can be captured by artificial multi-agent games (Kirby, 2002; Mordatch and Abbeel, 2018), in which agents have to communicate about some shared input space (e.g., images). A common *emergent communication* protocol has been adopted in a large body of recent research: a speaker encodes a piece of information into a sequence of discrete symbols (emergent language) and a listener then aims to decipher the sequence and recover the original piece of information (Lazaridou et al., 2017; Havrylov and Titov, 2017; Lazaridou et al., 2018; Bouchacourt and Baroni, 2018; Chaabouni et al., 2019; Li and Bowling, 2019; Chaabouni et al., 2020; Luna et al., 2020; Kharitonov and Baroni, 2020, *inter alia*).

The present work is partly inspired by the work of Lee et al. (2018), who train agents to communicate about images with their natural language captions and use their parameters as encoder-decoders for machine translation. However, this framework relies on the availability of natural language captions (whereas we use only *artificial* languages emerging from *raw* images). Moreover, it does not cast EC as pretraining followed by NMT few-shot fine-tuning; rather, it learns a model in a single stage. These differences make our approach not only applicable to truly resource-lean languages but also substantially superior in performance on a same dataset such as English-German Multi30k (see § 5).

Another strand of recent research (Lowe et al., 2019; Lowe et al., 2020; Lazaridou et al., 2020) aims at enhancing emergent communication success by encouraging agents to imitate natural language data supplied at the beginning of training. Our work goes the opposite direction and investigates whether an emergent communication protocol pretrained without any human language data can benefit downstream NLP applications such as machine translation.

Few-shot Neural Machine Translation. Our work addresses the problem of few-shot machine translation with limited parallel data. Differently from previous methods (Lample et al., 2018b; Lample et al., 2018a; Lample et al., 2018c; Gu et al., 2018a; Artetxe et al., 2018), our approach does not draw upon auxiliary language data for pretraining, which usually consists of machine translation tasks on other languages (Gu et al., 2018b) or domains (Sharaf et al., 2020), multilingual training (Aharoni et al., 2019; Liu et al., 2020), language model pretraining on monolingual data (Conneau and Lample, 2019; Siddhant et al., 2020), back-translation techniques on monolingual data (Platanios et al., 2018; Edunov et al., 2018), leveraging bilingual dictionaries (Duan et al., 2020), treebanks (Ponti et al., 2018), or image captions (Nakayama and Nishida, 2017; Elliott and Kádár, 2017; Lee et al., 2018).

On the contrary, we ground our neural model on visual knowledge acquired from agent interactions without any observation of human language, and then fine-tune our model on translation tasks even with as few as 500 to $1,000$ training instances. We rely on few-shot MT as a standard, well-known, and sound testbed to empirically validate the crucial question of this work, that is, whether emergent communication pretraining without any natural language data can inform models of language.

Inductive Biases for Language. Finally, a series of recent works has investigated how to construct neural models that are inductively biased towards learning new natural languages. This endeavour is motivated both by the need of sample efficiency and concerns of cognitive realism, as children can acquire language from limited stimuli (Chomsky, 1978). In particular, neural weights reflecting linguistic universals in phonotactics can be learned via approximate Bayesian inference (Ponti et al., 2019b) or meta-learning (McCoy et al., 2020). Papadimitriou and Jurafsky (2020) found that recurrent models pretrained on non-linguistic data with latent structure (such as music or code) facilitate natural language tasks.

To our knowledge, we are the first to propose grounded communication as a non-linguistic source for

pretraining, based on the hypothesis that modal and functional knowledge is a crucial inductive bias for fast and effective language acquisition.

3 Model Architecture

The proposed method comprises the standard two stages of transfer learning. First, as detailed in § 3.1, we pretrain *two speaker-listener agents* via emergent communication on image referential games. We then recombine[1] the pretrained EC agents to construct NMT encoder-decoder networks (see Figure 1), and *fine-tune* the networks on a small number of parallel sentence pairs, as we describe in § 3.2. At the fine-tuning stage, we also add an Adapter layer between the translation encoder and decoder, and further leverage two variants of regularisation with annealing, which are outlined in § 3.3. An illustrative overview of the proposed method is provided in Figure 1.

3.1 Emergent Communication Pretraining

EC pretraining consists in the following referential game: an image is seen only by a *speaker*, while a *listener* must guess the correct image among a set of distractors based on a message generated by the speaker. Cooperation and communication therefore arise due to information asymmetry between the two players. This setup follows previous work (Havrylov and Titov, 2017) with one core difference: like Lee et al. (2018), we train two agents, each consisting of a speaker and a listener, one for the source language, and another for the target language. Contrary to Lee et al. (2018), who rely on image-caption pairs for the supervised training of the speaker agents, we employ only unlabelled images to train communication protocols in an unsupervised way. The artificial language developed by agents is not explicitly constrained to resemble any natural language. We denote the two agents as $Agent_s = \{Speaker_s, Listener_s\}$ and $Agent_t = \{Speaker_t, Listener_t\}$. In our implementation, following recent work (Graesser et al., 2019; Resnick et al., 2020; Chaabouni et al., 2020; Kharitonov and Baroni, 2020; Lowe et al., 2020), both speakers and listeners are instantiated as single-layer Gated Recurrent Units (GRUs) (Chung et al., 2014).[2] The pretraining process of $Agent_s$ follows these steps:

Image Set Preparation. Let us denote the set of N images as $\mathcal{D}_I$. At each training step, an input image I_i, $i = 1, 2, ..., N$ and a set of K confounding images (i.e., negative examples) $C_i \subset \{I_j | I_j \in \mathcal{D}_I, j \neq i\}$, $|C_i| = K < N$ are randomly chosen from the entire set $\mathcal{D}_I$. Images are represented as 2,048-dimensional feature vectors extracted from a ResNet-50 CNN (He et al., 2016).

Message Generation. $Speaker_s$ takes the input image I_i and outputs a message $\mathbf{m}$ describing the image, a sequence of discrete symbols of variable length. The generation comes to a halt when the special end-of-sentence symbol is emitted or the maximum message length L_{max} is reached. Since $\mathbf{m}$ comprises discrete symbols, in order to make the model end-to-end differentiable, we adopt the Gumbel-Softmax distribution (Jang et al., 2017; Maddison et al., 2017) to draw samples from categorical distributions of emergent tokens while making the gradient flow.[3] The generation of the discrete symbol m_t at each time step t can be described by the following:

$$\mathbf{h}_t^s = \begin{cases} \text{GRU}_{Speaker}(<\text{bos}>, \text{MLP}_1(I_i)) & t = 0 \\ \text{GRU}_{Speaker}(m_{t-1}, \mathbf{h}_{t-1}^s) & t > 0 \end{cases} \quad (1)$$
$$m_t = \text{Gumbel-Softmax}(\text{MLP}_2(\mathbf{h}_t^s))$$

Here, $\mathbf{h}_t^s$ represents the hidden state at time step t, <bos> stands for the special beginning-of-sentence symbol, while MLP for multilayer perceptron. The parameters for MLP_1 are shared by both $Speaker$ and $Listener$ and map image features into input vectors for the GRU layer. A second MLP_2 is used by

[1] Note that we use each listener module as an MT encoder and each speaker module as a decoder. In addition, we train two separate agents because vocabulary sizes of SRC and TRG languages are different and we adopt disjoint input embeddings.

[2] We will experiment with Transformer-based architectures (Vaswani et al., 2017) in future work. Our choice of GRU is also partially motivated by recent results in few-shot MT showing on-par or even slightly stronger performance of recurrent networks over Transformers when only a small number of parallel sentences are available (Zhou et al., 2019).

[3] Another common approach is based on reinforcement learning, but recent work suggests that it is less effective and converges more slowly than Gumbel-Softmax for EC tasks (Havrylov and Titov, 2017; Lee et al., 2018).

$Speaker$ to project each GRU hidden state—one for each time step—into vectors with dimensionality equal to the predefined vocabulary size of the emergent language.

Image Inference. Given the input image, the generated message describing the image, and K confounding images, $Listener_s$ must now guess the correct input image among the distractors. To do so, a second GRU layer decodes the message generated by the $Speaker$ as follows:

$$\mathbf{h}_t^l = \begin{cases} \text{GRU}_{Listener}(m_0, \mathbf{0}) & t = 0 \\ \text{GRU}_{Listener}(m_t, \mathbf{h}_{t-1}^l) & t > 0 \end{cases} \tag{2}$$

$\mathbf{h}_t^l$ now denotes the $Listener$'s hidden state at time step t. The hidden state at the last time step, $\mathbf{h}_{|\mathbf{m}|}^l$, is used to reason over the correct image I_i and K distractors C_i, and guess which image is the one described by the $Speaker$. Given the message $\mathbf{m}$ and any image I we define a compatibility score based on the inverse squared error (Lee et al., 2018):

$$\text{score}(\mathbf{m}, I) = \left\| \mathbf{h}_{|\mathbf{m}|}^l - \text{MLP}_1(I) \right\|_2^{-2} \tag{3}$$

We then minimise the cross-entropy loss, treating the set of compatibility scores as logits, to optimise the agent parameters for the image referential game:

$$\mathcal{L} = -\mathbb{E}_{I_i \in \mathcal{D}_I} \mathbb{E}_{\mathbf{m}} \log \underbrace{\left(\frac{e^{\text{score}(\mathbf{m}, I_i)}}{\sum_{I_j \in \{I_i \cup C_i\}} e^{\text{score}(\mathbf{m}, I_j)}} \right)}_{P(\text{guess}=I_i | \mathbf{m}, I_i, C_i)} \tag{4}$$

In a nutshell, $Speaker_s$ takes an input from the image domain, then encodes it into a message in the emergent language domain. The message conveys information that has to be transferred back to the image domain by $Listener_s$ in order to solve the cooperative game. The same process is repeated alternating between $Agent_s$ and $Agent_t$.[4]

3.2 NMT Fine-Tuning and Adapters

After EC pretraining of $Agent_s$ and $Agent_t$, we recombine their speaker and listener modules into a standard sequence-to-sequence encoder-decoder neural machine translation (NMT) architecture, as shown in Figure 1. Let us denote a training set of *n* parallel sentences in the source and the target language as $\mathcal{D}$: $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$. The model then predicts the output sequence of the i-th parallel sentence:

$$P(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) = \prod_t P(y_t^{(i)} \mid \mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}) \tag{5}$$

In Eq. (5), $\mathbf{y}_{<t}^i$ represents the first $t-1$ tokens in the target language sentence $\mathbf{y}^i$, and the input sentence $\mathbf{x}^i$ is encoded as a fixed-length hidden vector by the encoder, following the standard sequence-to-sequence procedure (Sutskever et al., 2014). The sequence loss is defined as follows:

$$\mathcal{L}_{sequence} = -\mathbb{E}_{(\mathbf{x}^i, \mathbf{y}^i) \in \mathcal{D}} \log P(\mathbf{y}^i | \mathbf{x}^i) \tag{6}$$

The source-to-target translation model consists of $Listener_s$ (input: emergent language domain, output: image domain) and $Speaker_t$ (input: image domain, output: emergent language domain) and we denote their RNN parameters as $\mathbf{w}^\star$: these are transferred to MT fine-tuning. After fine-tuning on a small set of source-to-target sentence pairs, the model can perform the translation task. In an analogous manner, the target-to-source model is composed of $Listener_t$ and $Speaker_s$.

To compensate for the lack of an intermediate image domain in MT, at the fine-tuning stage we add an *Adapter module* in between encoders and decoders. Adapters are small neural modules that contain additional trainable parameters which facilitate quicker and more effective domain adaptation in computer

[4]We train $Agent_s$ and $Agent_t$ separately, as in preliminary experiments we found that sharing a global MLP_1 for image projection does not affect NMT performance.

vision (Rebuffi et al., 2017; Rebuffi et al., 2018) and, more recently, NLP tasks (Houlsby et al., 2019; Stickland and Murray, 2019; Pfeiffer et al., 2020b; Bapna and Firat, 2019; Pfeiffer et al., 2020a). A notable difference compared to prior work is that during the fine-tuning stage we train jointly both the Adapter and the model parameters (which are transferred from EC). Our Adapter modules follow a simple architecture from prior work (Houlsby et al., 2019), and comprise linear layers with residual connections and dropout, as illustrated in Figure 1.

3.3 Regularisation with Annealing

During fine-tuning, we also add to the objective an annealed regulariser for the encoder-decoder parameters (which, on the other hand, does not apply to the adapter module). These parameters are initialised using the parameters $\mathbf{w}^\star$ transferred from the EC agents. We can then define a regularisation term that prevents the parameters $\mathbf{w}$ from drifting away from their initialisation $\mathbf{w}^\star$ during fine-tuning (Duong et al., 2015):

$$\mathfrak{R} = \alpha \ \|\mathbf{w} - \mathbf{w}^\star\|^2 \tag{7}$$

where α is a positive real-valued tunable hyper-parameter denoting the strength of the regularisation penalty. Note that this amounts to placing a prior $\mathcal{N}(\mathbf{w}^\star, \mathbf{I}\alpha^{-1})$ on the encoder-decoder parameters. However, the contribution of the log-prior in Eq. (7) to the posterior probability of the parameters should stay fixed, whereas the contribution of the negative log-likelihood in Eq. (6) should grow linearly with the number of examples. In other words, the likelihood should be able to overwhelm the prior in the limit of infinite data. For this reason, the importance of the regulariser should be gradually decayed over fine-tuning steps. Therefore, we propose two variants of regularisation with annealing, labelled REG-A (exponential decay) and REG-B (inverse multiplicative decay). At the fine-tuning step k:

$$\text{REG-A: } \mathfrak{R}(k) = \alpha\lambda^k \ \|\mathbf{w} - \mathbf{w}^\star\|^2 \tag{8}$$

$$\text{REG-B: } \mathfrak{R}(k) = \frac{\alpha}{k} \ \|\mathbf{w} - \mathbf{w}^\star\|^2 \tag{9}$$

λ is a real-valued hyper-parameter from the interval $[0, 1)$ that controls the decay steepness. The final objective function is then as follows:

$$\min_{\mathbf{w}} \mathcal{L}_{sequence} + \mathfrak{R} \tag{10}$$

where $\mathcal{L}_{sequence}$ is provided by Eq. (6), and $\mathfrak{R}$ is one of REG-A or REG-B.

4 Experimental Setup

EC Pretraining is based on the MS COCO data set (Lin et al., 2014). We randomly select 50,000 images for training and 5,000 for validation.[5] For each image, a 2,048-dimensional feature vector is extracted from ResNet-50 (He et al., 2016). The input vocabulary size for EC is equal to the BPE vocabulary size during MT fine-tuning. However, since human language data are excluded, note that there is no alignment between EC and MT BPE vocabularies.[6] The maximum message length, L_{max}, is set to an integer around the average length, in terms of BPE tokens, of the MT training sets: 15-18 in our experiments. We do not impose additional constraints on the generated messages' length.[7] We later profile the impact of L_{max} on MT performance in §5.

The layer size is 256 for the input embeddings and 512 for the hidden layers. We use Adam (Kingma and Ba, 2015) with a learning rate of $lr = 0.001$. The dropout rate is set to 0.1 and the Gumbel-Softmax temperature is set to 1. The number of distracting images is $K = 255$ during training, and $K = 127$ in evaluation. Experiments on the validation set achieve the prediction accuracy of $> 99\%$ in all EC experimental runs, i.e., the listener is able to guess the single correct image from a set of 128 images almost always. We analyse the impact of EC prediction accuracy on few-shot MT performance later in §5.

[5] Again, we stress that we do not leverage image captions (available only for few languages in COCO) at all in our setup.

[6] The only exception is the end-of-sequence token <eos>.

[7] We only prevent the speakers from producing <eos> at the beginning of their output message. Without any constraints, the messages typically occupy the entire maximum allowed length L_{max}.

Machine Translation experiments are conducted on two standard datasets: Multi30k and Europarl. The Multi30k data set (Elliott et al., 2016; Barrault et al., 2018), originally devised for multi-modal MT, contains multilingual captions for $\approx 30k$ images. We discard images and run text-only fine-tuning and evaluation on English-German (EN-DE) and English-Czech (EN-CS) in both directions. We rely on the default training set of 29,000 pairs of parallel sentences, which we also subsample to simulate true few-shot scenarios: we randomly select 500, 1,000, and 10,000 sentence pairs for the lower-resource setups. In all experimental runs, we use the original validation set spanning 1,014 sentence pairs and the default test set spanning 1,000 pairs.

We also run experiments on Europarl data (Koehn, 2005) from OPUS (Tiedemann, 2009) for two language pairs: English-Romanian (EN-RO) and English-French (EN-FR), again in both directions. We retain only sentences with a length between 5 and 15 words to construct data sets whose average sentence length is similar to that of Multi30k. We then randomly sample 10,000 parallel sentences as our (largest) training set, while two other disjoint random samples of 1,500 sentence pairs are used for validation and test, respectively. As with Multi30k, we again sample 500 and 1,000 training instances from the full set of 10k examples to simulate few-shot settings.

For each language pair, we lowercase and tokenise the data using byte-pair encoding (BPE) (Sennrich et al., 2016). Our BPE vocabularies are derived from all 29,000 training pairs (for the Multi30k language pairs) and 10,000 training pairs (for the Europarl language pairs). We again use Adam in the same configuration as EC pretraining, except for setting the dropout rate to 0.2. The hyper-parameters of the annealed regulariser are set to $\alpha = 5$ and $\lambda = 0.998$ based on the scores on the EN-DE validation set (in the 1k training setup) and fixed to those values in all other experiments and for all other language pairs. For a fair comparison, the other hyper-parameters for fine-tuning are set identically to the NMT baseline introduced in the next paragraph.

NMT Baseline and Evaluation Details. The baseline NMT model is the standard seq2seq model whose architecture is exactly the same as our proposed model, but now with randomly initialised parameters (rather than transferred from EC). We extensively search the hyper-parameter space of the baseline model (Sennrich and Zhang, 2019) and adopt Adam optimiser with learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e$-08, a dropout rate of 0.2, a batch size of 128, a hidden-state size of 512, an embedding size of 256, and a max sequence length of 80. For all models, we rely on beam search with beam size 12 for decoding. The evaluation metric is BLEU-4 (Post, 2018).

5 Results and Analysis

In what follows, we report the NMT results of our proposed model on all language pairs. We then perform an ablation study highlighting the individual contributions—of the customised adapter layer, the strategies for annealing the regulariser, and emergent communication pretraining—to the final results. Finally, we assess the impact of the rate of communication success and maximum sequence length on downstream NMT performances.

Main Results. The BLEU scores of the model leveraging both EC pretraining and adapters are shown in Table 1 for the Multi30k dataset, and in Table 2 for Europarl. The results reveal sweeping gains on all language pairs and in both translation directions. These are especially accentuated in Czech and Romanian (e.g. +196.7% in EN-CS and +115.1% in RO-EN for 1k samples), which are arguably more distant from English than German and French. This suggests that our method might be particularly suited for languages that are the most challenging in real-world scenarios. Moreover, we note that the gains do not fade away as more training examples become available. For instance, while the relative improvements on the baseline decrease from +132.7% in the 500-shot setting to +29.0% in the 29k-shot setting (DE-EN), the absolute improvements remain consistent (+5.75 BLEU and +6.41 BLEU, respectively). Most importantly, the results clearly suggest the usefulness of EC pretraining on a downstream natural language task.

Ablation Study. In order to disentangle the contribution of each separate component of the full model, in Table 3 we report the results on two language pairs (EN-DE and RO-EN) for different combinations of the recurrent baseline, EC pretraining, adapters, and regulariser annealing strategies. We find that all the

	Model	0.5k Samples	1k Samples	10k Samples	29k Samples
EN-DE	Baseline	4.28	5.78	15.23	20.36
	EC Transferred + Adapter + REG-A	8.21	**10.77** $\uparrow^{86.3\%}$	19.93	23.99
	EC Transferred + Adapter + REG-B	**8.44** $\uparrow^{97.1\%}$	10.46	**21.59** $\uparrow^{41.7\%}$	**25.92** $\uparrow^{27.3\%}$
DE-EN	Baseline	4.33	6.41	15.92	22.09
	EC Transferred + Adapter + REG-A	**10.08**$\uparrow^{132.7\%}$	**12.81** $\uparrow^{99.8\%}$	20.31	25.65
	EC Transferred + Adapter + REG-B	10.04	12.13	**22.11** $\uparrow^{38.8\%}$	**28.50** $\uparrow^{29.0\%}$
EN-CS	Baseline	1.47	1.84	9.27	14.73
	EC Transferred + Adapter + REG-A	3.44	**5.46**$\uparrow^{196.7\%}$	13.33	17.58
	EC Transferred + Adapter + REG-B	**3.64**$\uparrow^{147.6\%}$	4.96	**13.62** $\uparrow^{46.9\%}$	**19.07** $\uparrow^{29.4\%}$
CS-EN	Baseline	5.71	6.69	15.15	19.94
	EC Transferred + Adapter + REG-A	**9.08** $\uparrow^{59.0\%}$	10.94	18.56	22.80
	EC Transferred + Adapter + REG-B	8.47	**11.05** $\uparrow^{65.1\%}$	**19.51** $\uparrow^{28.7\%}$	**25.29** $\uparrow^{26.8\%}$

Table 1: BLEU scores of the full model from §3 in the few-shot translation task on Multi30k with varying number of parallel sentences (N Samples). ↑ represents the highest score, associated with its relative gain over the baseline.

	Model	0.5k Samples	1k Samples	10k Samples
EN-RO	Baseline	1.71	2.83	7.37
	EC Transferred + Adapter + REG-A	3.58	**5.79**$\uparrow^{104.5\%}$	10.53
	EC Transferred + Adapter + REG-B	**3.71**$\uparrow^{116.9\%}$	5.62	**11.35**$\uparrow^{54.0\%}$
RO-EN	Baseline	1.96	3.03	9.15
	EC Transferred + Adapter + REG-A	**4.49**$\uparrow^{129.0\%}$	**6.52**$\uparrow^{115.1\%}$	12.03
	EC Transferred + Adapter + REG-B	4.43	6.10	**13.00** $\uparrow^{42.0\%}$
EN-FR	Baseline	1.95	2.50	6.42
	EC Transferred + Adapter + REG-A	2.96	4.52	8.81
	EC Transferred + Adapter + REG-B	**3.52** $\uparrow^{80.5\%}$	**4.63**$\uparrow^{85.2\%}$	**9.71**$\uparrow^{51.2\%}$
FR-EN	Baseline	1.83	2.40	6.64
	EC Transferred + Adapter + REG-A	3.28	**4.20**$\uparrow^{75.0\%}$	8.92
	EC Transferred + Adapter + REG-B	**3.64**$\uparrow^{98.9\%}$	4.12	**9.73**$\uparrow^{46.5\%}$

Table 2: BLEU scores of the full model from §3 in the few-shot translation task on Europarl with varying number of parallel sentences (N Samples). ↑ represents the highest score, associated with its relative gain over the baseline.

components improve the translation quality regardless of the amount of training data. Taking the case of EN-DE $0.5k$ as an example, the baseline achieves a BLEU of 4.28. On top of this, EC pretraining boosts this result to 6.48, adding the adapter layer to 5.21. When EC and adapters are combined, they yield a BLEU of 7.52.

Interestingly, the only finding in counter-tendency to this pattern is that the intersection of emergent communication pretraining and regulariser annealing decreases the performance compared with the baseline. Instead, when further combined with the adapters, it turns out to be the best configuration with 8.44 BLEU. This demonstrates that EC and the adapters work in synergy and play different roles, in retaining old knowledge and in acquiring novel information, respectively.

Lastly, by comparing the two strategies to anneal the regulariser during maximum-a-posteriori inference, we find no evidence favouring one or the other. Table 1, Table 2, and Table 3 show that while REG-A (exponential decay) achieves equal or better performance in 0.5k and 1k settings, REG-B (inverse multiplicative decay) shows its strength in 10k and 29k settings. A comparison on EN-DE between these two and a regulariser without annealing is shown in Appendix, where our regularisers gain an edge in all few-shot settings.[8]

Influence of EC Properties on MT Fine-Tuning. Finally, we investigate how the properties of the

[8]We also tried learning the prior diagonal variance via elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) but its performance is inferior to all other regularisers in our experiments, although it remains superior to the baseline.

	Model	0.5k Samples	1k Samples	10k Samples
EN-DE	Baseline	4.28	5.78	15.23
	Baseline + Adapter	5.21	7.25	16.90
	EC Transferred	6.48	8.47	16.33
	EC Transferred + REG-A	3.79 ↓	4.88 ↓	16.13
	EC Transferred + REG-B	4.17 ↓	5.72 ↓	16.60
	EC Transferred + Adapter	7.52	9.25	17.59
	EC Transferred + Adapter + REG-A	8.21	**10.77**↑$^{86.3\%}$	19.93
	EC Transferred + Adapter + REG-B	**8.44**↑$^{97.1\%}$	10.46	**21.59**↑$^{41.7\%}$
RO-EN	Baseline	1.96	3.03	9.15
	Baseline + Adapter	2.39	3.66	9.74
	EC Transferred	3.02	4.97	10.16
	EC Transferred + REG-A	1.57 ↓	2.12 ↓	7.20 ↓
	EC Transferred + REG-B	1.09 ↓	1.61 ↓	8.43 ↓
	EC Transferred + Adapter	**4.73**↑$^{141.3\%}$	6.11	11.74
	EC Transferred + Adapter + REG-A	4.49	**6.52**↑$^{115.1\%}$	12.03
	EC Transferred + Adapter + REG-B	4.43	6.10	**13.00**↑$^{42.0\%}$

Table 3: Ablation experiments. ↓ indicates the case when an added component reduces BLEU by at least 0.4 BLEU points; ↑ represents the highest score, associated with its relative gain over the main baseline.

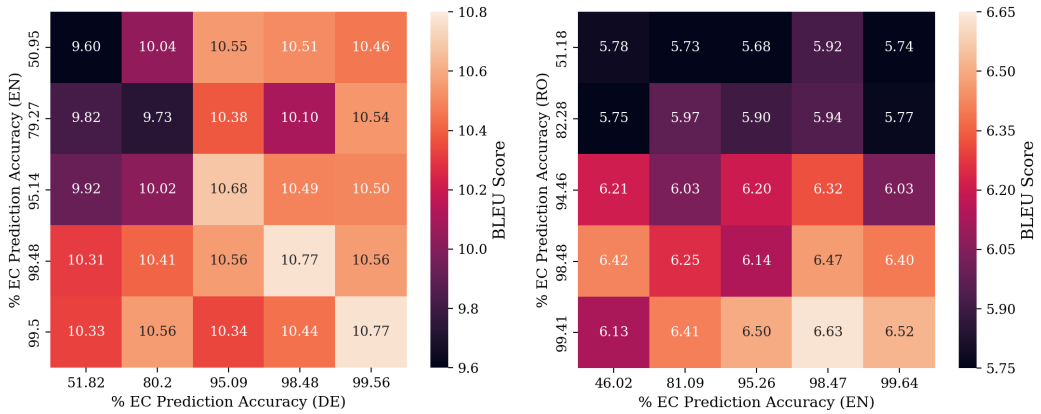

Figure 2: Impact of EC prediction accuracy on NMT BLEU scores for EN-DE (**left**) and RO-EN (**right**). All BLEU scores are obtained in the '1k Samples' setup with the full model variant EC Transferred + Adapter + REG-A.

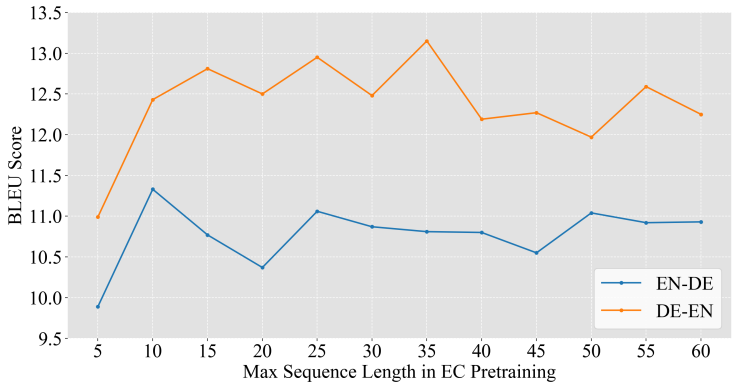

Figure 3: Impact of maximum EC message length (L_{max}) on NMT performance. All BLEU scores are obtained in the '1k Samples' setup with the full model variant EC Transferred + Adapter + REG-A.

artificial languages developed through EC affect downstream NMT performance. Most significantly, this can also be interpreted as a tool to evaluate whether emergent languages display affinities with natural languages. If this is the case, in fact, they should provide the correct inductive bias for NLP tasks and improve the sample efficiency of neural models.

First, we focus on the rate of communication success. During the EC pretraining stage, we save and

evaluate models every 50 training steps to pick up models with the desired level of accuracy. We run experiments on EN-DE and RO-EN (1k samples) and select five $Agent_s$ and $Agent_t$ whose prediction accuracy is near 50%, 80%, 95%, 98.5% and 99.5%, respectively. During fine-tuning we try all their 25 possible combinations. As shown in Figure 2, the prediction accuracy for $Agent_s$ and $Agent_t$ does positively correlate with MT performance. However, this trend is not strict and absolute, and sometimes sub-optimal EC models may fare better in the downstream task.

Second, we focus on the influence of maximum sequence length. In our main results, we have set L_{max} around the average sentence length (in BPE) of NMT training sets. However, we now show that this is not strictly necessary. In EN-DE and DE-EN, the average length for both languages and all splits (training, valid and test sets) are almost the same, 15. We vary L_{max} in steps of 5 from 5 to 60. In all these settings, we control for accuracy, only selecting models with a rate of communication success of $99.45 \pm 0.16\%$. The results are illustrated in Figure 3. They show that, with the exception of $L_{max} = 5$, for all the higher values MT performance is not particularly affected by L_{max}.

Further Discussion. One interesting question concerns what kind of knowledge exactly has been learned and transferred to the fine-tuning task. Of course, the pretrained EC model does not contain any information about either SRC or TRG languages. In fact, if the adapter is trained at the MT fine-tuning stage in isolation (freezing encoder and decoder to the initialisation values), MT performance turns out to be 0 in terms of BLEU score. What is more, it remains an open question whether the real-world grounding represented by image features plays a role in MT fine-tuning. If this were the case, one would expect higher gains in Multi30k than Europarl, as it consists of image captions. However, this does not occur in practice. As possible alternative hypotheses, EC pretraining might learn functional aspects of language (Wagner et al., 2003; Wittgenstein, 2009; Lazaridou et al., 2017; Lazaridou et al., 2020), i.e., the capability of agents to communicate and interact, or some language-universal structural properties, similar to Papadimitriou and Jurafsky (2020). We hope that future work will shed light on this matter.

We also note that without the adapter and the regulariser, the gains on MT are relatively limited. Hence, we must additionally stress the importance and synergy of both these modules to bridge between pretraining and fine-tuning tasks. On the one hand, initialisation and regularisation avoid catastrophic forgetting of old knowledge and drifts to parameter regions unfit for communication on referential games. On the other hand, the adapter module allows a drift from the image domain and thus results in fast adaptation to the new knowledge.

6 Conclusion and Future Work

We have demonstrated that an extreme pretraining paradigm without any human language data, but rather based on emergent communication (EC) in referential games, provides an inductive bias for learning natural languages. In theory, it makes this paradigm applicable to any of the world's languages, most of which suffer from the paucity of annotated data. In particular, we focused on neural machine translation (NMT) with limited parallel data as a downstream task. Our results across several language pairs and in different few-shot setups indicate that parameter initialisations informed by EC pretraining, in combination with adapter modules and annealed regularisation, yield higher accuracy and sample efficiency than baselines without any EC pretraining. Vice versa, we argued that NMT performance can also be interpreted as an extrinsic evaluation protocol for emergent communication models: it can assess to which extent emergent languages reflect properties found in natural languages. In particular, we discovered that communication success rate is well correlated with BLEU scores, whereas maximum sequence length is not impactful. In the future, we plan to experiment with other state-of-the-art NMT architectures, apply our method to extremely low-resource languages, and extend the scope of our work to other tasks beyond NMT.

7 Acknowledgements

We thank all the anonymous reviewers for their suggestions and comments. Our work is supported by the ERC Consolidator Grant LEXICAL (no 648909). EMP, IV, and AK are also funded through the Google Faculty Research Award 2018 for Natural Language Processing.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of NAACL-HLT 2019*, pages 3874–3884.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of ICLR 2018*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1538–1548.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Léon Bottou and Olivier Bousquet. 2008. The tradeoffs of large scale learning. In *Proceedings of NeurIPS 2008*, pages 161–168.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of EMNLP 2018*, pages 981–985.

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. In *Proceedings of NeurIPS 2020*.

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

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019. Word-order biases in deep-agent emergent communication. In *Proceedings of ACL 2019*, pages 5166–5175.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of ACL 2020*, pages 4427–4442.

Noam Chomsky. 1978. A naturalistic approach to language and cognition. *Cognition and Brain Theory*, 4(1):3–22.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS 2019*, pages 7057–7067.

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 2020*, pages 8440–8451.

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

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

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of ACL 2020*, pages 1570–1579.

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 2015*, pages 845–850.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2020. *Ethnologue: Languages of the World.* SIL International, xxiii edition.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP 2018*, pages 489–500.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of IJCNLP 2017*, pages 130–141.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Laura Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of EMNLP-IJCNLP 2019*, pages 3698–3708.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018a. Universal neural machine translation for extremely low resource languages. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of NAACL-HLT 2018*, pages 344–354.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of EMNLP 2018*, pages 3622–3631.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, pages 770–778.

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 2019*, pages 2790–2799.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018*, pages 328–339.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-softmax. In *Proceedings of ICLR 2017*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the ACL*, 8:64–77.

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

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

Eugene Kharitonov and Marco Baroni. 2020. Emergent language generalization and acquisition speed are not tied to compositionality. *arXiv preprint arXiv:2004.03420*.

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

Simon Kirby. 2002. Natural language from artificial life. *Artif. Life*, 8(2):185–215.

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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

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 2017*, pages 2962–2967.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR 2018*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *Proceedings of ICLR 2018*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP 2018*, pages 5039–5049.

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. In *Proceedings of EMNLP 2020*.

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

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *Proceedings of ICLR 2018*.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of ACL 2020*, pages 7663–7674.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2018. Emergent translation in multi-agent communication. In *Proceedings of ICLR 2018*.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV 2014*, pages 740–755.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2019. Learning to learn to communicate. In *Proceedings of the 1st Adaptive & Multitask Learning Workshop*.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. In *Proceedings of ICLR 2020*.

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. *arXiv preprint arXiv:2004.03868*.

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

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. In *Proceedings of CogSci 2020*.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of AAAI*, pages 1495–1502.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

Isabel Papadimitriou and Dan Jurafsky. 2020. Pretraining on non-linguistic structure as a tool for analyzing learning bias in language models. *arXiv preprint arXiv:2004.14601*.

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

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting Transformers. In *Proceedings of EMNLP 2020: System Demonstrations*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of EMNLP 2020*.

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

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

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, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. Towards zero-shot language modeling. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2893–2903.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

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. *CoRR*, abs/1910.10683.

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

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proceedings of NeurIPS 2017*, pages 506–516.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of CVPR 2018*, pages 8119–8127.

Cinjon Resnick, Abhinav Gupta, Jakob N. Foerster, Andrew M. Dai, and Kyunghyun Cho. 2020. Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of AAMAS 2020*, pages 1125–1133.

Ferdinand de Saussure. 1916. *Cours de linguistique générale, ed.* Payot. Edited by C. Bally and A. Sechehaye.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of ACL 2019*, pages 211–221.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725.

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot NMT adaptation. *arXiv preprint arXiv:2004.02745*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *CoRR*, abs/2005.04816.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of ICML 2019*, volume 97, pages 5926–5936.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of ICML 2019*, pages 5986–5995.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NeurIPS 2014*, pages 3104–3112.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of RANLP 2009*, pages 237–248.

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 2017*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Proceedings of NeurIPS 2016*, pages 3630–3638.

Kyle Wagner, James A Reggia, Juan Uriagereka, and Gerald S Wilkinson. 2003. Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR 2019*.

Ludwig Wittgenstein. 2009. *Philosophical investigations*. John Wiley & Sons.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP 2019*, pages 833–844.

Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig. 2019. Handling syntactic divergence in low-resource machine translation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1388–1394.

A Appendices

A.1 Regulariser without Annealing

In Table 4, we compare the BLEU scores of our proposed annealed regularisers (REG-A and REG-B) and a regulariser without annealing (NA) on EN-DE. ↓ indicates when the newly added module reduces the BLEU score by at least 0.4 BLEU points, and ↑ represents the highest gain compared with the baseline.

	Model	0.5k Samples	1k Samples	10k Samples	29k Samples
EN-DE	Baseline	4.28	5.78	15.23	20.36
	EC Transferred	6.48	8.47	16.33	21.43
	EC Transferred + NA, α = 5e-3	3.51 ↓	5.01 ↓	7.36 ↓	8.07 ↓
	EC Transferred + NA, α = 5e-4	5.42 ↓	6.94 ↓	16.69	21.84
	EC Transferred + NA, α = 5e-5	6.31	7.98 ↓	16.36	22.93
	EC Transferred + NA, α = 5e-6	6.8	9.35	16.35	21.89
	EC Transferred + NA, α = 5e-7	7.02	8.9	16.08	21.53
	EC Transferred + NA, α = 5e-8	6.82	8.39	16.59	20.88
	EC Transferred + REG-A	3.79 ↓	4.88 ↓	16.13	21.97
	EC Transferred + REG-B	4.17 ↓	5.72 ↓	16.60	23.19
	EC Transferred + Adapter	7.52	9.25	17.59	22.85
	EC Transferred + Adapter + NA, α = 5e-3	8.24	9.79	12.76 ↓	13.15 ↓
	EC Transferred + Adapter + NA, α = 5e-4	8.26	10.13	20.16	22.51
	EC Transferred + Adapter + NA, α = 5e-5	7.91	9.64	19.37	24.50
	EC Transferred + Adapter + NA, α = 5e-6	7.44	9.33	17.64	22.98
	EC Transferred + Adapter + REG-A	8.21	**10.77** $\uparrow^{86.3\%}$	19.93	23.99
	EC Transferred + Adapter + REG-B	**8.44** $\uparrow^{97.1\%}$	10.46	**21.59** $\uparrow^{41.7\%}$	**25.92** $\uparrow^{27.3\%}$

Table 4: Regularisers with and without annealing.