A. Complementary Theoretical Intuition for SIL and Its Limitation

We here provide a complementary intuition of Seeded Iterated Learning by referring to some mathematical tools that were used to study Iterated Learning dynamics in the general case. These are not the rigorous proof but guide the design of SIL.

One concern is that, since natural language is not fully compositional, whether iterated learning may favor the emergence of a new compositional language on top of the initial one. In this spirit, Griffiths & Kalish (2005); Kalish et al. (2007) modeled iterated learning as a Markov Process, and showed that vanilla iterated learning indeed converges to a language distribution that (i) is independent of the initial language distribution, (ii) depends on the student language before the inductive learning step.

Fortunately, Chazelle & Wang (2017) show iterated learning can converge towards a distribution close to the initial one with high probability if the intermediate student distributions remain close enough of their teacher distributions and if the number of training observations increases logarithmically with the number of iterations.

This theoretical result motivates one difference between our framework and classical iterated learning: as we want to preserve the pretrained language distribution, we do not initialize the new students from scratch as in (Li & Bowling, 2019; Guo et al., 2019; Ren et al., 2020) because the latter approach exert a uniform prior on the space of language, while we would like to add a prior that favors natural language (e.g. favoring language whose token frequency satisfies Zipf's Law).

A straightforward instantiation of the above theoretic results is to initialize new students as the pretrained model. However we empirically observe that, periodically resetting the model to initial pretrained model would quickly saturate the task score. As a result, we just keep using the students from the last imitation learning for the beginning of new generation, as well as retain the natural language properties from pretraining checkpoint.

However, we would also point out the limitation of existing theoretical results in the context of deep learning: The theoretical iterated learning results assume the agent to be perfect Bayesian learner (e.g. Learning is infering the posterior distribution of hypothesis given data). However, we only apply standard deep learning training procedure in our setup, which might not have this property. Because of the assumption of perfect Bayesian learner, (Chazelle & Wang, 2019) suggests to use training sessions with increasing length. However in practice, increasing k_2 may be counter-productive because of overfitting issues (especially when we have limited number of training scenarios).

B. Lewis Game

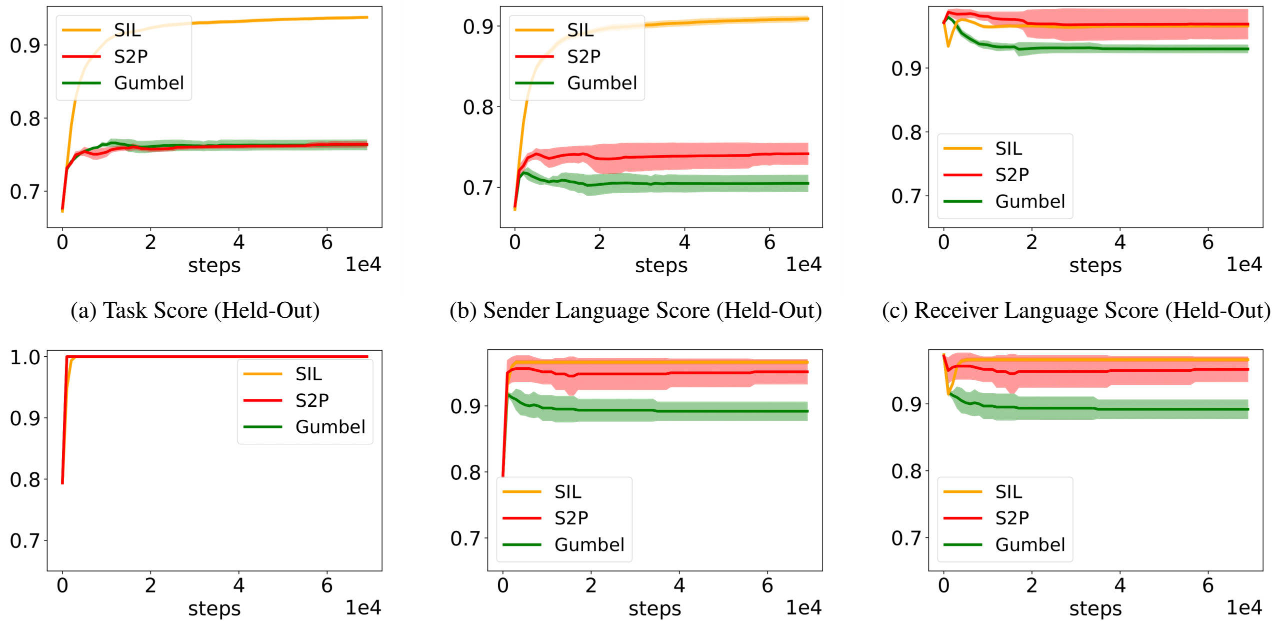
B.1. Experiment Details

In the Lewis game, the sender and the receiver architecture are 2-layer MLP with a hidden size of 200 and no-activation ($ReLU$ activations lead to similar scores). During interaction learning, we use a learning rate of 1e-4 for SIL. We use a learning rate of 1e-3 for the baselines as it provides better performance on the language and score tasks. In both cases, we use a training batch size of 100. For the teacher imitation phase, the student uses a learning rate of 1e-4.

In the Lewis game setting, we generate objects with $p = 5$ properties, where each property may take $t = 5$ values. Thus, it exists 3125 objects, which we split into 3 datasets: the pretraining, the interactive, and testing datasets. The pretraining split only contains 10 combination of objects. As soon as we provide additional objects, the sender and receiver fully solve the game by using the target language, which is not suitable to study the language drift phenomenon. The interactive split contains 30 objects. This choice is arbitrary, and choosing a additional objects gives similar results. Finally, the 3.1k remaining objects are held-out for evaluation.

B.2. Additional Plots

We sweep over different Gumbel temperatures to assess the impact of exploration on language drift. We show the results with Gumbel temperature $\tau = 1, 10$ in Fig 13 and Fig 12. We observe that the baselines are very sensitive to Gumbel temperature: high temperature both decreases the language and tasks score. On the other side, Seeded Iterated Learning perform equally well on both temperatures and manage to maintain both task and language accuracies even with high temperature.



(a) Task Score (Held-Out)

(b) Sender Language Score (Held-Out)

(c) Receiver Language Score (Held-Out)

(d) Task Score (Train)

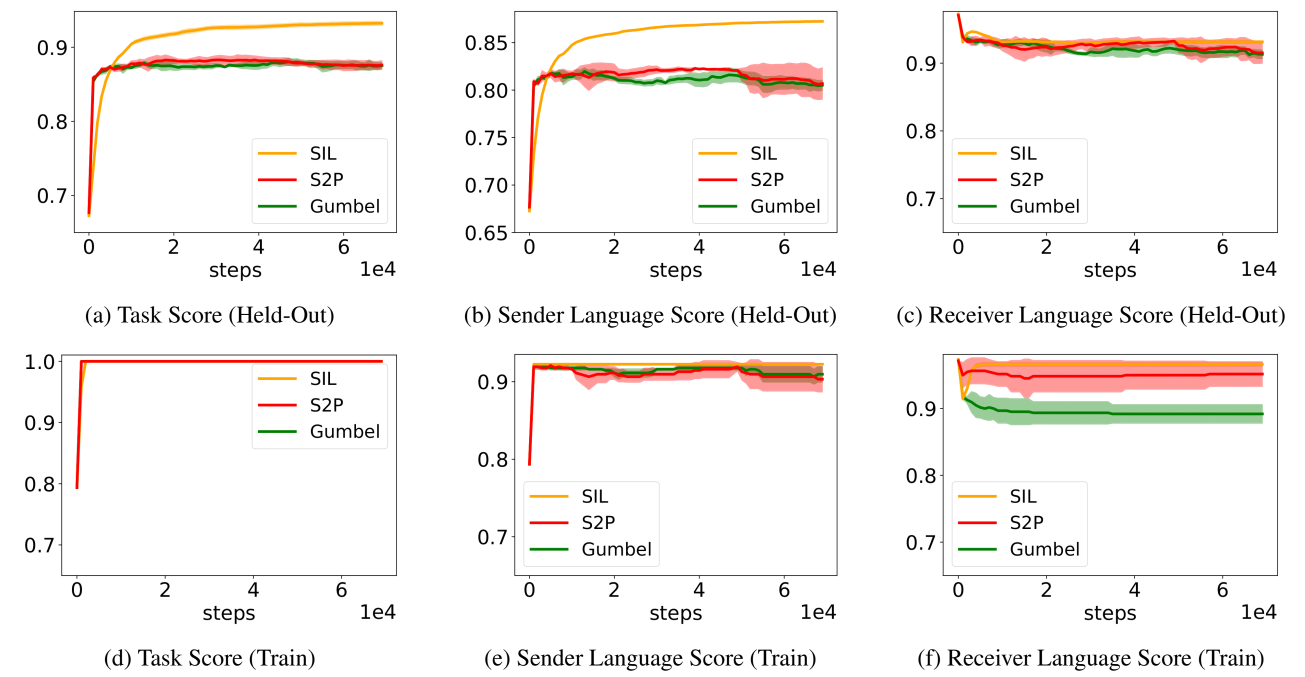
(e) Sender Language Score (Train)

(f) Receiver Language Score (Train)

Figure 12. Complete training curves for Task score and sender grounding in Lewis Game comparing SIL vs baselines for $\tau = 10$ on the held-out dataset (bottom), and the interactive training split (bottom). We observe that the three methods reach 100% accuracy on the training task score, but their score differs on the held-out split. For SIL we use $k_1 = 1000, k_2 = k_2' = 400$.

B.3. Tracking Language Drift with Token Accuracy

To further visualize the language drift in Lewis game, we focus on the evolution of on the probability of speaking different word when facing the same concept. Formally, we track the change of conditional probability $s(w|c)$. The result is in Figure 14.



(a) Task Score (Held-Out) (b) Sender Language Score (Held-Out) (c) Receiver Language Score (Held-Out)

(d) Task Score (Train) (e) Sender Language Score (Train) (f) Receiver Language Score (Train)

Figure 13. Complete training curves for Task score and sender grounding in Lewis Game comparing SIL vs baselines for $\tau = 1$ on the held-out dataset (bottom), and the interactive training split (bottom). For SIL we use $k_1 = 1000, k_2 = k_2' = 400$.

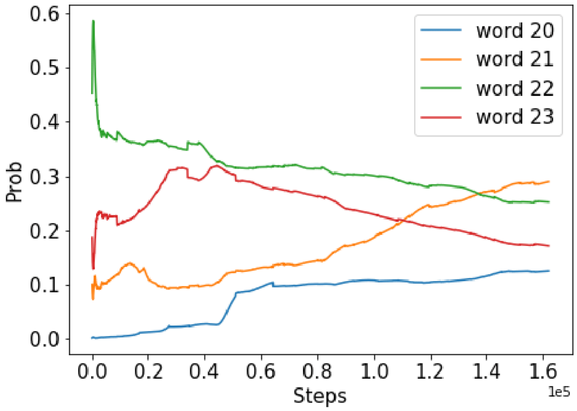


Figure 14. Change of conditional probability $s(w|c)$ where $c = 22$ and $w = 20, 21, 22, 23$. Following pretraining, $s(22|22)$ start with the highest probability. However, language drift gradually happens and eventually word 21 replaces the correct word 22.

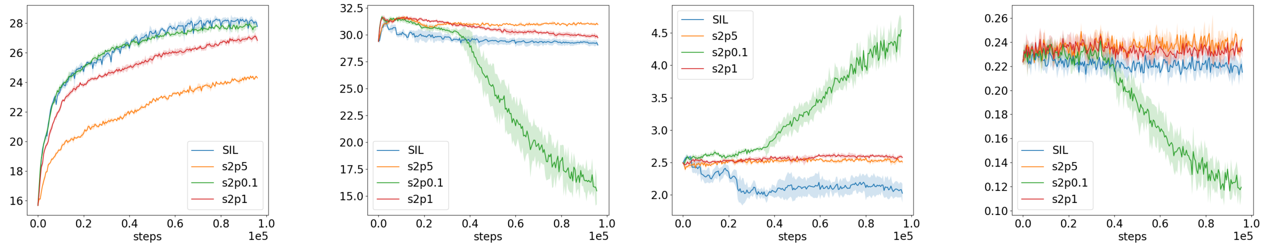
C. Translation Game

C.1. Data Preprocessing

We use Moses to tokenize the text (Koehn et al., 2007) and we learn byte-pair-encoding (Sennrich et al., 2016) from Multi30K (Elliott et al., 2016) with all language. Then we apply the same BPE to different dataset. Our vocab size for En, Fr, De is 11552, 13331, and 12124.

Table 2. Translation Game Results. The checkmark in "ref len" means the method use reference length to constrain the output during training/testing. ↑ means higher the better and vice versa. Our results are averaged over 5 seeds, and reported values are extracted for the best BLEU(De) score during training. We here use a Gumbel temperature of 0.5.

Method		ref len	BLEU↑		NLL↓	R1%↑
			De	En		
Lee et al. (2019)	Pretrained	N/A	16.3	27.18	N/A	N/A
	PG	✓	24.51	12.38	N/A	N/A
	PG+LM+G	✓	28.08	24.75	N/A	N/A
Ours	Pretrained	N/A	15.68	29.39	2.49	21.9
	Fix Sender	N/A	22.02 ± 0.18	29.39	2.49	21.9
	Gumbel		27.11 ± 0.14	14.5 ± 0.83	5.33 ± 0.39	9.7 ± 1.2
	Gumbel	✓	26.94 ± 0.20	23.41± 0.50	5.04 ± 0.01	18.9 ± 0.8
	S2P($\alpha = 0.1$)		27.43± 0.36	19.16 ± 0.63	4.05 ± 0.16	13.6 ± 0.7
	S2P($\alpha = 1$)		27.35± 0.19	29.73 ± 0.15	2.59 ± 0.02	**23.7 ± 0.7**
	S2P($\alpha = 5$)		24.64± 0.16	**30.84 ± 0.07**	2.51 ± 0.02	23.5 ± 0.5
	NIL		**28.29± 0.16**	29.4 ± 0.25	**2.15 ± 0.12**	21.7 ± 0.2



(a) BLEU De (Task Score) (b) BLEU En (c) NLL (d) R1

Figure 15. S2P has a trade-off between the task score and the language score while SIL is consistently high with both metrics.

C.2. Model Details and Hyperparameters

The model is a standard seq2seq translation model with attention (Bahdanau et al., 2015). Both encoder and decoder have a single-layer GRU (Cho et al., 2014) with hidden size 256. The embedding size is 256. There is a dropout after embedding layers for both encoder and decoder For decoder at each step, we concatenate the input and the attention context from last step.

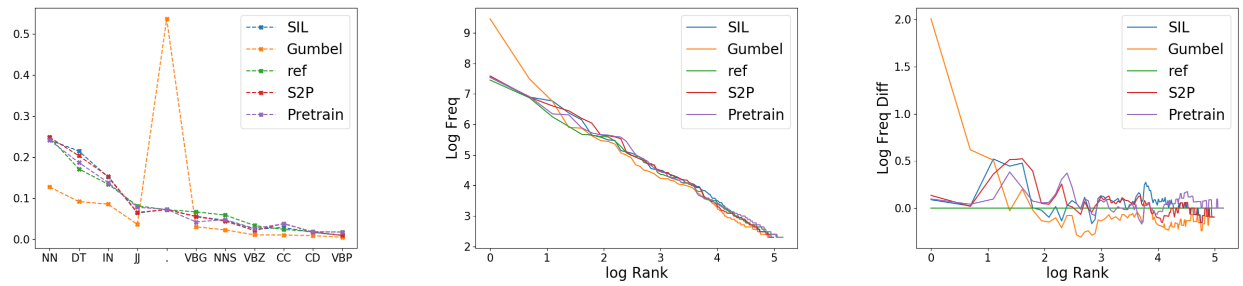
Pretraining For Fr-En agent, we use dropout ratio 0.2, batch size 2000 and learning rate 3e-4. We employ a linear learning rate schedule with the anneal steps of 500k. The minimum learning rate is 1e-5. We use Adam optimizer (Kingma & Ba, 2014) with $\beta = (0.9, 0.98)$. We employ a gradient clipping of 0.1. For En-De, the dropout ratio is 0.3. We obtain a BLEU score of 32.17 for Fr-En, and 20.2 for En-De on the IWSLT test dataset (Cettolo et al., 2012).

Finetuning During finetuning, we use batch size 1024 and learning rate 1e-5 with no schedule. The maximum decoding length is 40 and minimum decoding length is 3. For iterated learning, we use $k_1 = 4000$, $k_2 = 200$ and $k'_2 = 300$. We set Gumbel temperature to be 5. We use greedy sample from teacher speaker for imitation.

C.3. Language Model and Image Ranker Details

Our language model is a single-layer LSTM (Hochreiter & Schmidhuber, 1997) with hidden size 512 and embedding size 512. We use Adam and learning rate of 3e-4. We use a batch size of 256 and a linear schedule with 30k anneal steps. The language model is trained with captions from MSCOCO (Lin et al., 2014). For the image ranker, we use the pretrained ResNet-152 (He et al., 2016) to extract the image features. We use a GRU (Cho et al., 2014) with hidden size 1024 and embedding size 300. We use a batch size of 256 and use VSE loss (Faghri et al., 2017). We use Adam with learning rate of 3e-4 and a schedule with 3000 anneal steps (Kingma & Ba, 2014).

C.4. Language Statistics



(a) POS tag distribution.

(b) Word Frequency Analysis

(c) Difference of Log of Word Frequency

Figure 16. Language statistics on samples from different method.

We here compute several linguistic statistics on the generated samples to assess language quality.

POS Tag Distribution We compute the Part-of-Speech Tag (POS Tag (Marcus et al., 1993)) distribution by counting the frequency of POS tags and normalize it. The POS tag are sorted according to their frequencies in the reference, and we pick the 11 most common POS tag for visualization, which are:

- NN Noun, singular or mass
- DT Determiner
- IN Preposition or subordinating conjunction
- JJ Adjective
- VBG Verb, gerund or present participle
- NNS Noun, plural
- VBZ Verb, 3rd person singular present
- CC Coordinating conjunction
- CD Cardinal number

The results are shown in Figure 16a. The peak on “period” show that Gumbel has tendency of repeating periods at the end of sentences. However, we observe that both S2P and

Word Frequency For each generated text, we sort the frequency of the words and plot the log of frequency vs. log of rank. We set a minimum frequency of 50 to exclude long tail results. The result is in Figure 16b.

Word Frequency Difference To further visualize the difference between generated samples and reference, we plot the difference between their log of word frequencies in Figure 16c.

S2P, Reward Shaping and KL Minimization We find that multiple baselines for countering language drift can be summarized under the framework of KL minimization. Suppose the distribution of our model is P and the reference model is Q. Then in order to prevent the drift of P, we minimize $KL(P|Q)$ or $KL(Q|P)$ in addition to normal interactive training. We show that $KL(P|Q)$ is related to the reward shaping Lee et al. (2019) and $KL(Q|P)$ is related to S2P Gupta et al. (2019).

One find that

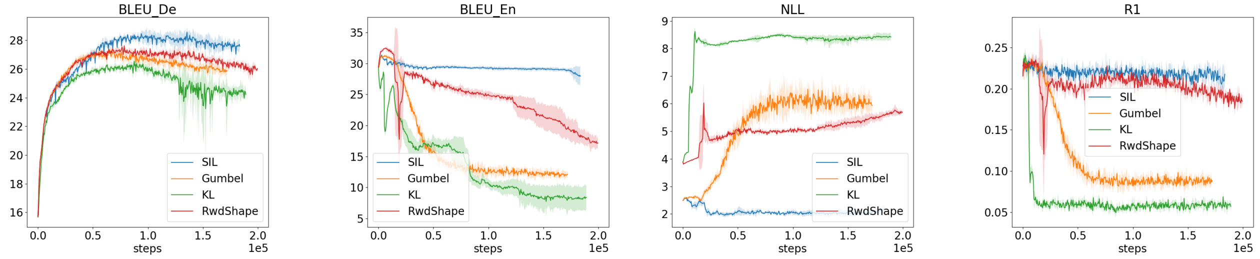
$$\min KL(Q|P) = \min E_Q[\log Q - \log P] = \max H(Q) + E_Q[\log P] = \max E_Q[\log P]$$

We can find that S2P can be obtained if we let Q to be the underlying data distribution. In the same spirit, one find that

$$\min KL(P|Q) = \max H(P) + E_P[\log Q]$$

The first term is equivalent to an entropy regularization term, while the second term is maximizing the reward $\log Q$. We implement the baseline $KL(P|Q)$ by using the same Gumbel Softmax trick to optimize the term $E_P[\log Q]$, where Q is the pretrained language model from MSCOCO captions. The training loss is defined as $\mathcal{L} = \mathcal{L}_{selfplay} + \beta\mathcal{L}_{kl}$. We only show $\beta = 0.1$ here and other values of β do not yield better result.

The result can be found in Figure 17. Since KL can be decomposed into a reward reshaping term and a entropy maximizing term. So I compare to an extra baseline RwdShaping which remove the entropy term since encouraging exploration would make the drift worse. We find that KL baseline is even worse than Gumbel baseline for both task score and language score, mainly due to its emphasis on entropy maximization term. By removing that term, we see RwdShape can outperform Gumbel on both task score and language score, but compared with SIL, RwdShape still has larger drift.



(a) BLEU De (Task Score) (b) BLEU En (c) NLL (d) R1

Figure 17. Comparison between SIL and different KL baselines

D. Human Evaluation

We here assess whether our language drift evaluation correlates with human judgement. To do so, we performed a human evaluation with two pairwise comparison tasks.

- In Task1, the participant picks the best English semantic translation while observing the French sentence.
- In Task2, the participant picks the best English translation from two candidates.

Thus, the participants are likely to rank captions mainly by their syntax/grammar quality in Task2, whereas they would also consider semantics in Task1, allowing us to partially disentangle structural and semantic drift.

For each task, we use the validation data from Multi30K (1013 French captions) and generate 4 English sentences for each French caption from the Pretrain, Gumbel, S2P, and SIL. We also retrieved the ground-truth human English caption. We then build the test by randomly sampling two out of five English captions. We gathered 22 people, and we collect about 638 pairwise comparisons for Task2 and 315 pairwise comparisons for Task1. We present the result in Table 4 and Table 5. I also include the binomial statistical test result where the null hypothesis is *methods are the same*, and the alternative hypothesis is *one method is better than the other one*.

Unsurprisingly, we observe that the Human samples are always preferred over generated sentences. Similarly, Gumbel is substantially less preferred than other models in both settings.

In Task 1(French provided), human users always preferred S2P and SIL over pretrained models with a higher win ratio. Oh the other hand when French is not provided, the human users prefer the pretrain models over S2P and SIL. We argue that while the pretrained model keeps generating gramartically correct sentences, its translation effectiveness is worse than both S2P and SIL since these two models go through the interactive learning to adapt to new domain.

Finally, SIL seems to be preferred over S2P by a small margin in both tasks. However, our current ranking is not conclusive, since we can see the significance level of comparisons among Pretrain, S2P, and SIL is not smaller enough to reject null hypothesis, especially in task 1 where we have less data points. In the future we plan to have a larger scale human evaluation to further differentiate these methods.

Table 3. The Win-Ratio Results. The number in row X and column Y is the empiric ratio that method X beats method Y according collected human pairwise preferences. We perform a naive ranking by the row-sum of win-ratios of each method. We also provide the corresponding P-values under each table. The null hypothesis is *two methods are the same*, while the alternative hypothesis is *two methods are different.*

Table 4. With French Sentences

	Gumbel	Pretrain	S2P	SIL	Human
Gumbel	0	0.25	0.15	0.12	0
Pretrain	0.75	0	0.4	0.4	0.13
S2P	0.84	0.6	0	0.38	0.21
SIL	0.88	0.6	0.63	0	0.22
Human	1	0.87	0.79	0.77	0
Ranking	Human(3.4), SIL(2.3), S2P(2.0), Pretrain(1.7), Gumbel(0.5)				
			P-values		
	Gumbel	Pretrain	S2P	SIL	Human
Gumbel	-	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$
Pretrain	$< 10^{-2}$	-	0.18	0.21	$< 10^{-2}$
S2P	$< 10^{-2}$	0.18	-	0.15	$< 10^{-2}$
SIL	$< 10^{-2}$	0.21	0.15	-	$< 10^{-2}$
Human	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	-

Table 5. Without French Sentences

	Gumbel	Pretrain	S2P	SIL	Human
Gumbel	0	0.16	0.12	0.13	0.02
Pretrain	0.84	0	0.69	0.59	0.15
S2P	0.88	0.31	0	0.38	0.05
SIL	0.86	0.41	0.62	0	0.01
Human	0.98	0.85	0.95	0.98	0
Ranking	Human(3.8), Pretrain(2.3), SIL(1.9), S2P(1.6), Gumbel(0.4)				
			P-values		
	Gumbel	Pretrain	S2P	SIL	Human
Gumbel	-	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$
Pretrain	$< 10^{-2}$	-	$< 10^{-2}$	0.08	$< 10^{-2}$
S2P	$< 10^{-2}$	$< 10^{-2}$	-	0.06	$< 10^{-2}$
SIL	$< 10^{-2}$	0.08	0.06	-	$< 10^{-2}$
Human	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	-

E. Samples

We list more samples from the Multi30k dataset with different baselines, i.e., Pretrain, Gumbel, S2P($\alpha = 1$. The Gumbel temperature is set to 0.5. The complete samples can be found in our code.

ref : a female playing a song on her violin .
Pretrain: a woman playing a piece on her violin .
Gumbel : a woman playing a piece on his violin . . . .
S2P : a woman playing a piece on his violin .
SIL : a woman playing a piece on his violin .

ref : a cute baby is smiling at another child .
Pretrain: a nice baby smiles at another child .
Gumbel : a nice baby smiles of another child . . . .

S2P : a nice baby smiles at another child .
SIL : a beautiful baby smiles smiles at another child .

ref : a man drives an old-fashioned red race car .
Pretrain: a man conducted an old race car .
Gumbel : a man drives a old race of red race . . . .
S2P : a man drives an old of the red race .
SIL : a man drives a old race of the red race .

ref : a man in a harness climbing a rock wall
Pretrain: a man named after a rock man .
Gumbel : a man ththththththththdeacdeaacc. of th. . . . . . .
S2P : a man ’s being a kind of a kind of a kind .
SIL : a man that the datawall of the datad.

ref : a man and woman fishing at the beach .
Pretrain: a man and a woman is a woman .
Gumbel : a man and a woman thaccbeach the beach . . . .
S2P : a man and a woman is in the beach .
SIL : a man and a woman that ’s going to the beach .

ref : a man cooking burgers on a black grill .
Pretrain: a man making the meets on a black slick of a black slick .
Gumbel : a man doing it of on a black barbecue . . . .
S2P : a man doing the kind on a black barbecue .
SIL : a man doing the datadon a black barbecue .

ref : little boy in cami crawling on brown floor
Pretrain: a little boy in combination with brown soil .
Gumbel : a small boy combincombinaccon a brown floor . . . brown . . . .
S2P : a small boy combining the kind of brown floor .
SIL : a small boy in the combination of on a brown floor .

ref : dog in plants crouches to look at camera .
Pretrain: a dog in the middle of plants are coming to look at the goal .
Gumbel : a dog in middle of of of of thlooking at looking at objeobje. . . . .
S2P : a dog in the middle of the plants to watch objective .
SIL : a dog at the middle of plants are going to look at the objective .

ref : men wearing blue uniforms sit on a bus .
Pretrain: men wearing black uniforms are sitting in a bus .
Gumbel : men wearing blue uniforms sitting in a bus . . . .
S2P : men wearing blue uniforms sitting in a bus .
SIL : men wearing blue uniforms are sitting in a bus .

ref : a group of scottish officers doing a demonstration .
Pretrain: a scottish officers group is doing a demonstration .
Gumbel : a group of officers scottish doing a dedemonstration . . . .
S2P : a group of officers scottish doing a demonstration .
SIL : a group of officers scottish doing a demo .

ref : the brown dog is wearing a black collar .
Pretrain: the brown dog is wearing a black collar .
Gumbel : the brown dog carries a black collar . . . .
S2P : the brown dog carries a black collar .
SIL : the brown dog is wearing a black collar .

ref : twp children dig holes in the dirt .
Pretrain: two children are going to dig holes in the earth .

Gumbel : two children dig holes in the planplanplanplan. . . . .
S2P : two children are going holes in the dirt .
SIL : two children dig holes in the earth .

ref : the skiers are in front of the lodge .
Pretrain: the health are in front of the bed .
Gumbel : the ththare ahead the thth. . . . . . .
S2P : the health are front of the whole .
SIL : the dataare are ahead of the datad.

ref : a seated man is working with his hands .
Pretrain: a man sitting working with his hands .
Gumbel : a man sitting working with his hands . . . .
S2P : a man sitting working with his hands .
SIL : a man sitting working with its hands .

ref : a young girl is swimming in a pool .
Pretrain: a girl swimming in a swimming pool .
Gumbel : a young girl swimming in a pool . . . .
S2P : a young girl swimming in a pool .
SIL : a young girl swimming in a pool .

ref : a small blond girl is holding a sandwich .
Pretrain: a little girl who is a sandwich .
Gumbel : a yedegirl holding a sandwich . . . .
S2P : a small 1girl holding a sandwich .
SIL : a small 1girl holding a sandwich .

ref : two women look out at many houses below .
Pretrain: two women are looking at many of the houses in the computer .
Gumbel : two women looking many of many houses in itdeacede. . . . .
S2P : two women looking at many houses in the kind .
SIL : two women looking at many houses in the data.

ref : a person is hang gliding in the ocean .
Pretrain: (wind up instead of making a little bit of the board) a person who is the board of the sailing .
Gumbel : (cdthinplace of acacc) a person does thacthof th-acin the ocean . . . .
S2P : (wind ’s instead of a kind) a person does the kind in the ocean .
SIL : (datadinstead of the input of the clinability) a person does the board in the ocean .

ref : a man in a green jacket is smiling .
Pretrain: a green man in the green man .
Gumbel : a man jacket green smiles . . . .
S2P : a man in jacket green smiles .
SIL : a man in the green jacket smiles .

ref : a young girl standing in a grassy field .
Pretrain: a girl standing in a meadow .
Gumbel : a young girl standing in a gmeadow . . . .
S2P : a young girl standing in a meadow .
SIL : a young girl standing in a meadow .