Grounded PCFG Induction with Images

Lifeng Jin and **William Schuler**
Department of Linguistics
The Ohio State University, Columbus, OH, USA
{jin, schuler}@ling.osu.edu

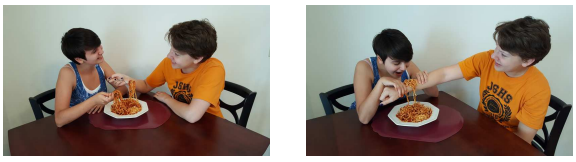
Abstract

Recent work in unsupervised parsing has tried to incorporate visual information into learning, but results suggest that these models need linguistic bias to compete against models that only rely on text. This work proposes grammar induction models which use visual information from images for labeled parsing, and achieve state-of-the-art results on grounded grammar induction on several languages. Results indicate that visual information is especially helpful in languages where high frequency words are more broadly distributed. Comparison between models with and without visual information shows that the grounded models are able to use visual information for proposing noun phrases, gathering useful information from images for unknown words, and achieving better performance at prepositional phrase attachment prediction.[1]

1 Introduction

Recent grammar induction models are able to produce accurate grammars and labeled parses with raw text only (Jin et al., 2018b, 2019; Kim et al., 2019b,a; Drozdov et al., 2019), providing evidence against the poverty of the stimulus argument (Chomsky, 1965), and showing that many linguistic distinctions like lexical and phrasal categories can be directly induced from raw text statistics. However, as computational-level models of human syntax acquisition, they lack semantic, pragmatic and environmental information which human learners seem to use (Gleitman, 1990; Pinker and MacWhinney, 1987; Tomasello, 2003).

This paper proposes novel grounded neural-network-based models of grammar induction which take into account information extracted from images in learning. Performance comparisons show that the proposed models achieve state-of-the-art results on multilingual induction datasets, even without help from linguistic knowledge or pretrained image encoders. Experiments show several specific benefits attributable to the use of visual information in induction. First, as a proxy to semantics, the co-occurrences between objects in images and referring words and expressions, such as the word *spaghetti* and the plate of spaghetti in Figure 1,[2] provide clues to the induction model about the syntactic categories of such linguistic units, which may complement distributional cues from word collocation which normal grammar inducers rely on solely for induction. Also, pictures may help disambiguate different syntactic relations: induction models are not able to resolve many prepositional phrase attachment ambiguities with only text — for example in Figure 1, there is little information in the text of *Mary eats spaghetti with a friend* for the induction models to induce a high attachment structure where *a friend* is a companion — and images may provide information to resolve these ambiguities. Finally, images may provide grounding information for unknown words when their syntactic properties are not clearly indicated by sentential context.



(a) friend as companion (b) friend as condiment

Figure 1: Examples of disambiguating information provided by images for the prepositional phrase attachment of the sentence *Mary eats spaghetti with a friend* (Gokcen et al., 2018).

[1]The system implementation and translated datasets used in this work can be found at https://github.com/lifengjin/imagepcfg.

[2]https://github.com/ajdagokcen/madlyambiguous-repo

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 396–408

2 Related work

Existing unsupervised PCFG inducers exploit naturally-occurring cognitive and developmental constraints, such as punctuation as a proxy to prosody (Seginer, 2007), human memory constraints (Noji and Johnson, 2016; Shain et al., 2016; Jin et al., 2018b), and morphology (Jin and Schuler, 2019), to regulate the posterior of grammars which are known to be extremely multimodal (Johnson et al., 2007). Models in Shi et al. (2019) also match embeddings of word spans to encoded images to induce unlabeled hierarchical structures with a concreteness measure (Hill et al., 2014; Hill and Korhonen, 2014). Additionally, visual information is observed to provide grounding for words describing concrete objects, helping to identify and categorize such words. This hypothesis is termed 'noun bias' in language acquisition (Gentner, 1982, 2006; Waxman et al., 2013), through which the early acquisition of nouns is attributed to nouns referring to observable objects. However, the models in Shi et al. (2019) also rely on language-specific branching bias to outperform other text-based models, and images are encoded by pretrained object classifiers trained with large datasets, with no ablation to show the benefit of visual information for unsupervised parsing. Visual information has also been used for joint training of prepositional phrase attachment models (Christie et al., 2016) suggesting that visual information may contain semantic information to help disambiguate prepositional phrase attachment.

3 Grounded Grammar Induction Model

The full grounded grammar induction model used in these experiments, ImagePCFG, consists of two parts: a word-based PCFG induction model and a vision model, as shown in Figure 2. The two parts have their own objective functions. The PCFG induction model, called NoImagePCFG when trained by itself, can be trained by maximizing the marginal probability $\mathsf{P}(\sigma)$ of sentences σ. This part functions similarly to previously proposed PCFG induction models (Jin et al., 2018a; Kim et al., 2019a) where a PCFG is induced through maximization of the data likelihood of the training corpus marginalized over latent syntactic trees.

The image encoder-decoder network in the vision model is trained to reconstruct the original image after passing through an information bottleneck. The latent encoding from the image encoder may be seen as a compressed representation of visual information in the image, some of which is semantic, relating to objects in the image. We hypothesize that semantic information can be helpful in syntax induction, potentially through helping three tasks mentioned above.

In contrast to the full model where the encoded visual representations are trained from scratch, the ImagePrePCFG model uses image embeddings encoded by pretrained image classifiers with parameters fixed during induction training. We hypothesize that pretrained image classifiers may provide useful information about objects in an image, but for grammar induction it is better to allow the inducer to decide which kind of information may help induction.

The two parts are connected through a syntactic-visual loss function connecting a syntactic sentence embedding projected from word embeddings and an image embedding. We hypothesize that visual information in the encoded images may help constrain the search space of syntactic embeddings of words with supporting evidence of lexical attributes such as concreteness for nouns or correlating adjectives with properties of objects.[3]

3.1 Induction model

The PCFG induction model is factored into three submodels: a nonterminal expansion model, a terminal expansion model and a split model, which distinguishes terminal and nonterminal expansions. The binary-branching non-terminal expansion rule probabilities,[4] and unary-branching terminal expansion rule probabilities in a factored Chomsky-normal-form PCFG can be parameterized with these three submodels. Given a tree as a set τ of nodes η undergoing non-terminal expansions $c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}$ (where $\eta \in \{1, 2\}^*$ is a Gorn address specifying a path of left or right branches from the root), and a set τ' of nodes η undergoing terminal expansions $c_\eta \rightarrow w_\eta$ (where w_η is the word at node η) in a parse of sentence σ, the marginal

[3] The syntactic nature of word embeddings indicates that any lexical-specific semantic information in these embeddings may be abstract, which is generally not sufficient for visual reconstruction. Experiments with syntactic embeddings show that it is difficult to extract semantic information from them and present visually.

[4] These include the expansion rules generating the top node in the tree.

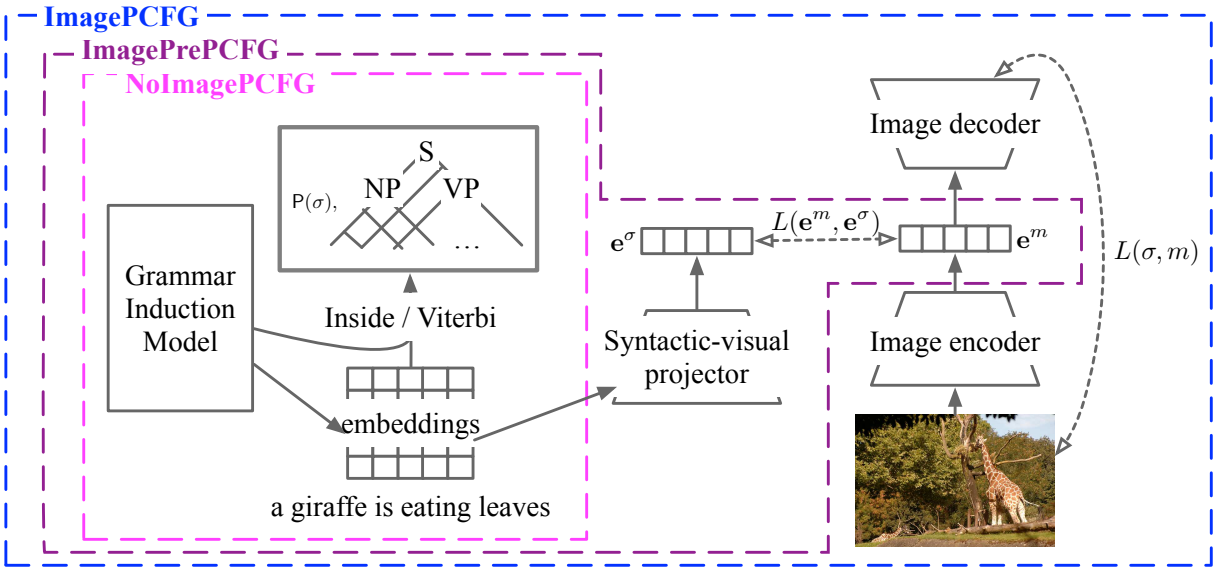


Figure 2: Different configurations of PCFG induction models: the model without vision (NoImagePCFG), the model with a pretrained image encoder (ImagePrePCFG) and the model with images (ImagePCFG.)

probability of σ can be computed as:

$$\mathsf{P}(\sigma) = \sum_{\tau,\tau'} \prod_{\eta\in\tau} \mathsf{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta\in\tau'} \mathsf{P}(c_\eta \rightarrow w_\eta) \tag{1}$$

We first define a set of Bernoulli distributions that distribute probability mass between terminal and nonterminal rules, so that the lexical expansion model can be tied to the image model (see Section 4.2):

$$\mathsf{P}(\text{Term} \mid c_\eta) = \underset{\{0,1\}}{\text{softmax}}(\text{ReLU}(\mathbf{W}_{\text{spl}}\, \mathbf{x}_{B,c_\eta} + \mathbf{b}_{\text{spl}})), \tag{2}$$

where c_η is a non-terminal category, $\mathbf{W}_{\text{spl}} \in \mathbb{R}^{2\times h}$ and $\mathbf{b}_{\text{spl}} \in \mathbb{R}^2$ are model parameters for hidden vectors of size h, and $\mathbf{x}_{B,c_\eta} \in \mathbb{R}^h$ the result of a multilayered residual network (Kim et al., 2019a). The residual network consists of B architecturally identical residual blocks. For an input vector $\mathbf{x}_{b-1,c}$ each residual block b performs the following computation:

$$\mathbf{x}_{b,c} = \text{ReLU}(\mathbf{W}_b\ \text{ReLU}(\mathbf{W}'_b\, \mathbf{x}_{b-1,c} + \mathbf{b}'_b) + \mathbf{b}_b) + \mathbf{x}_{b-1,c}, \tag{3}$$

with base case:

$$\mathbf{x}_{0,c} = \text{ReLU}(\mathbf{W}_0\, \mathbf{E}\, \delta_c + \mathbf{b}_0) \tag{4}$$

where δ_c is a Kronecker delta function – a vector with value one at index c and zeros everywhere else – and $\mathbf{E} \in \mathbb{R}^{d\times C}$ is an embedding matrix for each nonterminal category c with embedding size d, and $\mathbf{W}_0 \in \mathbb{R}^{h\times d}$, $\mathbf{W}_b, \mathbf{W}'_b \in \mathbb{R}^{h\times h}$ and $\mathbf{b}_0, \mathbf{b}_b, \mathbf{b}'_b \in \mathbb{R}^h$ are model parameters with latent representations of size h. B is set to 2 in all models following Kim et al. (2019a). Binary-branching non-terminal expansion rule probabilities for each non-terminal category c_η and left and right children $c_{\eta 1}\ c_{\eta 2}$ are defined as:

$$\mathsf{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}) = \mathsf{P}(\text{Term=0} \mid c_\eta) \cdot \underset{c_{\eta 1}, c_{\eta 2}}{\text{softmax}}(\mathbf{W}_{\text{nont}}\, \mathbf{E}\, \delta_{c_\eta} + \mathbf{b}_{\text{nont}}), \tag{5}$$

where $\mathbf{W}_{\text{nont}} \in \mathbb{R}^{C^2\times d}$ and $\mathbf{b}_{\text{nont}} \in \mathbb{R}^{C^2}$ are parameters of the model.

The lexical unary-expansion rule probabilities for a preterminal category c_η and a word w_η at node η are defined as:

$$\mathsf{P}(c_\eta \rightarrow w_\eta) = \mathsf{P}(\text{Term=1} \mid c_\eta) \cdot \frac{\exp(n_{c_\eta, w_\eta})}{\sum_w \exp(n_{c_\eta, w})} \tag{6}$$

$$n_{c,w} = \text{ReLU}(\mathbf{w}_{\text{lex}}^\top\, \mathbf{n}_{B,c,w} + b_{\text{lex}}) \tag{7}$$

where w is the generated word type, and $\mathbf{w}_{\text{lex}} \in \mathbb{R}^h$ and $b_{\text{lex}} \in \mathbb{R}$ are model parameters. Similarly,

$$\mathbf{n}_{b,c,w} = \text{ReLU}(\mathbf{W}''_b\ \text{ReLU}(\mathbf{W}'''_b\, \mathbf{n}_{b-1,c,w} + \mathbf{b}'''_b) + \mathbf{b}''_b) + \mathbf{n}_{b-1,c,w}, \tag{8}$$

with base case:

$$\mathbf{n}_{0,c,w} = \text{ReLU}(\mathbf{W}'_0 \begin{bmatrix} \mathbf{E}\, \delta_c \\ \mathbf{L}\, \delta_w \end{bmatrix}) + \mathbf{b}'_0) \tag{9}$$

where $\mathbf{W}'_0 \in \mathbb{R}^{h\times 2d}$, $\mathbf{W}''_b, \mathbf{W}'''_b \in \mathbb{R}^{h\times h}$ and $\mathbf{b}_0, \mathbf{b}''_b, \mathbf{b}'''_b \in \mathbb{R}^h$ are model parameters for latent representations of size h. $\mathbf{L}$ is a matrix of syntactic word embeddings for all words in vocabulary.

4 Vision model

The vision model consists of an image encoder-decoder network and a syntactic-visual projector. The image encoder-decoder network encodes an image into an image embedding and then decodes that back into the original image. This reconstruction constrains the information in the image embedding to be closely representative of the original image. The syntactic-visual projector projects word embeddings used in the calculation of lexical expansion probabilities into the space of image embeddings, building a connection between the space of syntactic information and the space of visual information.

4.1 The image encoder-decoder network

The image encoder employs a ResNet18 architecture (He et al., 2016) which encodes an image with 3 channels into a single vector. The encoder consists of four blocks of residual convolutional networks. The image decoder decodes an image from a visual vector generated by the image encoder. The image decoder used in the joint model is the image generator from DCGAN (Radford et al., 2016), where a series of transposed convolutions and batch normalizations attempts to recover an image from an image embedding.[5]

4.2 The syntactic-visual projector

The projector model is a CNN-based neural network which takes a concatenated sentence embedding matrix $\mathbf{M}^\sigma \in \mathbb{R}^{|\sigma|\times d}$ as input, where embeddings in $\mathbf{M}^\sigma$ are taken from $\mathbf{L}$, and returns the syntactic-visual embedding $\mathbf{e}^\sigma$. The jth full length-wise convolutional kernel is defined as a matrix $\mathbf{K}_j \in \mathbb{R}^{u_j\times k_j d}$ which slides across the sentence matrix $\mathbf{M}$ to produce a feature map, where u_j is the number of channels in the kernel, k_j is the width of the kernel, and d is the height of the kernel which is equal to the size of the syntactic word embeddings. Because the kernel is as high as the embeddings, it produces one vector of length u_j for each window. The full feature map $\mathbf{F}_j \in \mathbb{R}^{u_j\times H_j}$, where H_j is total number of valid submatrices for the kernel, is:

$$\mathbf{F}_j = \sum_h (\mathbf{K}_j \,\mathrm{vec}(\mathbf{M}^\sigma{}_{[h..k_j+h-1,*]}) + \mathbf{b}_j)\,\delta_h^\top. \quad (10)$$

Finally, an average pooling layer and a linear transform are applied to feature maps from different kernels:

$$\hat{\mathbf{f}} = [\mathrm{mean}(\mathbf{F}_1)\ \ldots\ \mathrm{mean}(\mathbf{F}_j)]^\top, \quad (11)$$

$$\mathbf{e}^\sigma = \tanh(\mathbf{W}_{\mathrm{pool}}\mathrm{ReLU}(\hat{\mathbf{f}}) + \mathbf{b}_{\mathrm{pool}}). \quad (12)$$

All $\mathbf{K}$s, $\mathbf{b}$s and $\mathbf{W}$s here are parameters of the projector.

5 Optimization

There are three different kinds of objectives used in the optimization of the full grounded induction model. The first loss is the marginal likelihood loss for the PCFG induction model described in Equation 1, which can be calculated with the Inside algorithm. The second loss is the syntactic-visual loss. Given the encoded image embedding $\mathbf{e}^m$ and the projected syntactic-visual embedding $\mathbf{e}^\sigma$ of a sentence σ, the syntactic-visual loss is the mean squared error of these two embeddings:

$$L(\mathbf{e}^m, \mathbf{e}^\sigma) = (\mathbf{e}^m - \mathbf{e}^\sigma)^\top(\mathbf{e}^m - \mathbf{e}^\sigma). \quad (13)$$

The third loss is the reconstruction loss of the image. Given the original image represented as a vector $\mathbf{i}^m$ and the reconstructed image $\hat{\mathbf{i}}^m$, the reconstruction objective is the mean squared error of the corresponding pixel values of the two images:

$$L(m) = (\mathbf{i}^m - \hat{\mathbf{i}}^m)^\top(\mathbf{i}^m - \hat{\mathbf{i}}^m). \quad (14)$$

Models with different sets of input optimize the three losses differently for clean ablation. NoImagePCFG, which learns from text only, optimizes the negative marginal likelihood loss (the negative of Equation 1) using gradient descent. The model with pretrained image encoders, ImagePrePCFG, optimizes the negative marginal likelihood and the syntactic-visual loss (Equation 13) simultaneously. The full grounded grammar induction model ImagePCFG learns from text and images jointly by minimizing all three objectives: negative marginal likelihood, syntactic-visual loss and image reconstruction loss (Equation 14):

$$L(\sigma, m) = -\mathsf{P}(\sigma) + L(\mathbf{e}^m, \mathbf{e}^\sigma) + L(m). \quad (15)$$

[5] Details of these models can be found in the cited work and the appendix.

6 Experiment methods

Experiments described in this paper use the MSCOCO caption data set (Lin et al., 2015) and the Multi30k dataset (Elliott et al., 2016), which contains pairs of images and descriptions of images written by human annotators. Captions in the MSCOCO data set are in English, whereas captions in the Multi30k dataset are in English, German and French. Captions are automatically parsed (Kitaev and Klein, 2018) to generate a version of the reference set with constituency trees.[6] In addition to these datasets with captions generated by human annotators, we automatically translate the English captions into Chinese, Polish and Korean using Google Translate,[7] and parse the resulting translations into constituency trees, which are then used in experiments to probe the interactions between visual information and grammar induction.

Results from models proposed in this paper — NoImagePCFG, ImagePrePCFG and ImagePCFG — are compared with published results from Shi et al. (2019), which include PRPN (Shen et al., 2018), ON-LSTM (Shen et al., 2019) as well as the grounded VG-NSL models which uses either head final bias (VG-NSL$_{+H}$) or head final bias and Fasttext embeddings (VG-NSL$_{+H+F}$) as inductive biases from external sources. All of these models only induce unlabeled structures and have been evaluated with unlabeled F1 scores. We additionally report the labeled evaluation score Recall-Homogeneity (Rosenberg and Hirschberg, 2007; Jin and Schuler, 2020) for better comparison between the proposed models. All evaluation is done on Viterbi parse trees of the test set from 5 different runs. Details about hyper-parameters and results on development data sets can be found in the appendix. However, importantly, the tuned hyperparameters for the grammar induction model are the same across the three proposed models, which facilitates direct comparisons among these models to determine the effect of visual information on induction.

6.1 Standard set: no replication of effect for visual information

Both unlabeled and labeled evaluation results are shown in Table 1 with left- and right-branching baselines. First, trees induced by the PCFG induction models are more accurate than trees induced with all other models, showing that the family of PCFG induction models is better at capturing syntactic regularities and provides a much stronger baseline for grammar induction. Second, using the NoImagePCFG model as a baseline, results from both the ImagePCFG model, where raw images are used as input, and the ImagePrePCFG model, where images encoded by pretrained image classifiers are used as input, do not show strong indication of benefits of visual information in induction. The baseline NoImagePCFG outperforms other models by significant margins on all languages in unlabeled evaluation. Compared to seemingly large gains between text-based models like PRPN and ON-LSTM[8] and the grounded models like VG-NSL$_{+H}$ on French and German observed by Shi et al. (2019), the only positive gain between NoImagePCFG and ImagePCFG shown in Table 1 is the labeled evaluation on French where ImagePCFG outperforms NoImagePCFG by a small margin. Because the only difference between NoImagePCFG and ImagePCFG models is whether the visual information influences the syntactic word embeddings, the results indicate that on these languages, visual information does not seem to help induction. The gain seen in previous results may therefore be from external inductive biases. Finally, the ImagePrePCFG model performs at slightly lower accuracies than the ImagePCFG model consistently across different languages, datasets and evaluation metrics, showing that the information needed in grammar induction from images is not the same as information needed for image classification, and such information can be extracted from images without annotated image classification data.

6.2 Languages with wider distribution of high-frequency word types: positive effect

One potential advantage of using visual information in induction is to ground nouns and noun phrases. For example, if images like in Figure 1 are consistently presented to models with sentences describing *spaghetti*, the models may learn the categorize words and phrases which could be linked with objects in images as nominal units and then bootstrap other lexical categories. However, in the test languages above, a narrow set of very high fre-

[6] The multilingual parsing accuracy for all languages used in this work has been validated in Fried et al. (2019) and verified in Shi et al. (2019).

[7] https://translate.google.com/.

[8] PCFG induction models where a grammar is induced generally perform better in parsing evaluation than sequence models where only syntactic structures are induced (Kim et al., 2019a; Jin et al., 2019).

Models	MSCOCO		Multi30k					
	English**		English**		German**		French**	
	F1	RH	F1	RH	F1	RH	F1	RH
Left-branching	23.3	-	22.6	-	34.7	-	19.0	-
Right-branching	21.4	-	11.3	-	12.1	-	11.0	-
PRPN	52.5±2.6	-	30.8±17.9	-	31.5±8.9	-	27.5±7.0	-
ON-LSTM	45.5±3.3	-	38.7±12.7	-	34.9±12.3	-	27.7±5.6	-
VG-NSL+H	53.3±0.2	-	38.7±0.2	-	38.3±0.2	-	38.1±0.6	-
VG-NSL+H+F	54.4±0.4	-	-	-	-	-	-	-
NoImagePCFG	**60.0**±8.2	**47.6**±10.0	**59.4**±7.7	**51.6**±8.5	**48.1**±5.2	**53.7**±5.2	**44.3**±5.1	43.8±5.2
ImagePrePCFG	55.6±7.5	42.3±7.3	47.0±7.0	40.5±7.2	46.2±7.4	51.1±8.0	42.6±10.3	43.4±10.8
ImagePCFG	55.1±2.7	42.5±1.5	48.2±4.9	40.5±5.0	47.0±5.5	51.8±8.4	43.6±5.5	**44.5**±6.3

Table 1: Averages and standard deviations of labeled Recall-Homogeneity and unlabeled F1 scores of various unsupervised grammar inducers on the MSCOCO and Multi30k caption datasets. VG-NSL+H: VG-NSL system with head final bias. VG-NSL+H+F: VG-NSL system with head final bias and Fasttext word embeddings.(** : the unlabeled performance difference between NoImagePCFG and ImagePCFG is significant $p < 0.01$.)

quency words such as determiners provide strong identifying information for nouns and noun phrases, which may greatly diminish the advantage contributed by visual information. In such cases, visual information may even be harmful, as models may attend to other information in images which is irrelevant to induction.

Korean, Polish and Chinese are chosen as representatives of languages with no definite articles, and in which statistical information provided by high frequency words is less reliable because there are more such word types. Table 2 shows the performance scores of the three proposed systems on these languages. Comparing to results in Table 1, the models with visual information in the input significantly outperform the baseline model, NoImagePCFG, on a majority of the additional test datasets. Figure 3 shows the correlation between the RH difference between the ImagePCFG model and the NoImagePCFG model on each language in an image dataset, and the distribution of high frequency words in that language, defined as the number of word types needed to account for 10% of the number of word tokens in the Universal Dependency (Nivre et al., 2016) corpus of a language.[9] The figure shows that the largest gain brought by visual information in induction is on Korean, where the number of high frequency word types is also highest. Results on Chinese and Polish

[9]Korean has 41, Chinese and Polish have 5, German has 4, English has 3 and French has 2.

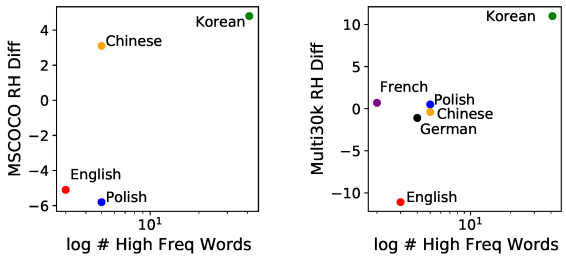


Figure 3: The correlation between number of word types needed to account for 10% of word tokens in a language (log # High Freq Words) and the RH gain from NoImagePCFG to ImagePCFG on different languages on the two different image datasets.

also show a benefit for visual information, although the gain is much smaller and less consistent. It also shows that when there is a trend of positive correlation between the number of high frequency words and the gain brought by visual information, factors other than high frequency words are at play as well in determining the final induction outcome for each dataset in each language in the visually grounded setup, which are left for investigation in future work.

7 Analysis of advantages of visual information

We hypothesize three specific ways that visual information may help grammar induction. First, a strong correlation between words and objects in images can help identification and categorization

Models on MSCOCO	Korean**		Polish**		Chinese**	
	F1	RH	F1	RH	F1	RH
NoImagePCFG	38.1±8.5	22.3±6.8	58.9±3.7	47.1±3.8	61.2±3.5	48.5±3.7
ImagePrePCFG	39.0±4.1	23.5±3.2	**60.5**±1.8	**49.8**±3.3	60.0±4.6	47.2±4.5
ImagePCFG	**45.0**±2.2	**27.1**±2.6	53.6±8.3	41.3±7.8	**64.9**±6.6	**51.2**±8.6
Models on Multi30k	**Korean****		**Polish**		**Chinese****	
	F1	RH	F1	RH	F1	RH
NoImagePCFG	30.7±5.6	22.8±3.1	49.6±4.6	39.9±5.1	**59.1**±3.3	**53.2**±4.7
ImagePrePCFG	27.1±4.4	19.9±3.4	48.4±3.1	38.3±2.9	57.9±7.0	51.0±7.7
ImagePCFG	**44.9**±1.3	**33.8**±2.1	**49.7**±7.2	**40.4**±6.1	58.5±3.2	52.8±4.6

Table 2: Averages and standard deviations of labeled Recall-Homogeneity and unlabeled F1 scores of various unsupervised grammar inducers on the MSCOCO and Multi30k caption datasets in the additional languages with high numbers of high-frequency word types. (** : the unlabeled performance difference between NoImagePCFG and ImagePCFG is significant $p < 0.01$.)

of nouns and noun phrases, especially on languages where nouns and noun phrases are not readily identifiable by neighboring high frequency words. Second, visual information may provide bottom-up information for unknown word embeddings. Languages where neighboring words can reliably predict the grammatical category of an unknown word may build robust representations of unknown word embeddings, but the construction of the UNK embedding may also benefit from bottom-up information from images, especially when sentential context is not enough to build informative UNK embeddings. Finally, semantic information inside images may be helpful in solving syntactic ambiguities like prepositional phrase attachment in languages like English. Results from experiments described below with the ImagePCFG and NoImagePCFG models show evidence of all three ways.

7.1 Grounding of nouns and noun phrases

The 'Noun bias' hypothesis (Gentner, 1982) postulates that visual information in the induction process may impact how words are categorized grammatically, and nouns may receive an advantage because they correspond to objects in images. However, objects in images are often described with phrases, not single words. For example, captions like *a red car is parked on the street*, are common in both caption datasets, where the objects in the image may associate more strongly with modifier words like *red* than the head noun *car*.

Evaluations are carried out on the parsed sentences of all languages from two caption datasets using a part-of-speech homogeneity metric (Rosenberg and Hirschberg, 2007) for measuring the part-of-speech accuracy, and an unlabeled NP recall score for measuring how many noun phrases in gold annotation are also found in the induced trees. Results in Figure 4 first show that the POS homogeneity scores from different models on the same induction dataset are extremely close to each other. Given that nouns are one of the categories with the most numerous tokens, the almost identical performance of POS homogeneity across different models indicates that the unsupervised clustering accuracy for nouns across different models is also very close, in contrast to substantial RH score differences on English and Korean.

However, NP recall scores show a pattern of performance ranking that resembles the ranking observed in Tables 1 and 2. For all datasets except for the Polish Multi30k dataset, when the RH score of ImagePCFG is higher than NoImagePCFG, the NP recall score for the ImagePCFG model is also higher. Significance testing with permutation sampling shows that all performance differences are significant ($p < 0.01$).[10] High accuracy on noun phrases is crucial to high accuracy of other constituents such as prepositional phrases and verb phrases, which usually contain noun phrases, and eventually leads to high overall accuracy. This result suggests that the benefit contributed by visual information works at phrasal levels, most likely

[10] Significance testing is not done on POS homogeneity due to the possibility that the same induced POS label may mean different things in different induced grammars.

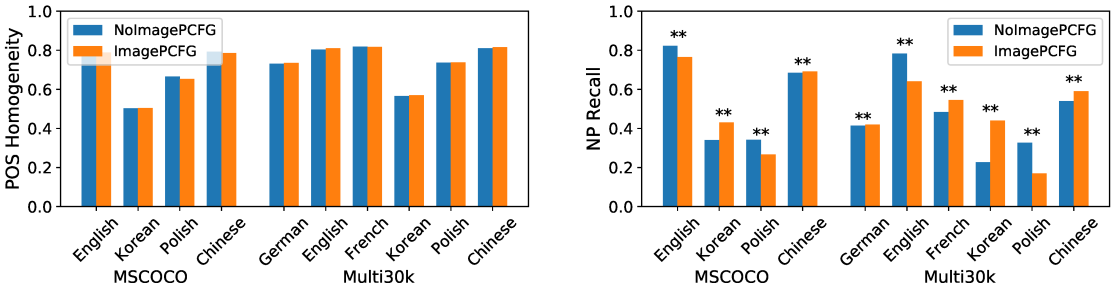


Figure 4: The POS Homogeneity and NP Recall scores for the ImagePCFG and NoImagePCFG models across the test languages (** : $p < 0.01$).

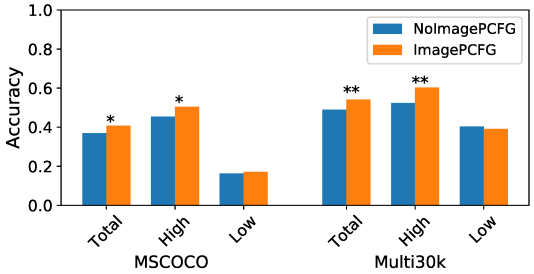


Figure 5: The average overall accuracy as well as accuracies for high and low attachment sentences in PP attachment evaluation for models with and without visual information (** : $p < 0.01$, * :$p < 0.05$).

grounding phrases, not words, with objects in images.

7.2 Informativeness of the UNK embedding

The informativeness of unknown word embeddings is tested among the induction models across different languages. An UNK test set is created by randomly replacing one word in one sentence with an UNK symbol if the sentence has no unknown words present. Table 3 shows the labeled evaluation results on the multilingual datasets.[11] First, performance on the UNK test sets on all languages is lower than on the normal test sets, showing that replacing random words with UNK symbols does impact performance. The performance ranking of the models on a majority of the languages is consistent with the ranking on the normal test set. The ranking of the models on one dataset, the Chinese Multi30k, is reversed on the UNK test set, where the ImagePCFG models show significantly higher performance than the NoImagePCFG models (Chinese: $p < 0.01$, permutation test on unlabeled F1). This result indicates that the ImagePCFG model in which visual information is supplied during training may have built more informative embeddings for the unknown word symbols, helping the model to outperform the model without visual information on a majority of datasets where UNK symbols are frequent.

[11]The unlabeled evaluation results can be found in the appendix.

7.3 Prepositional phrase attachment

Finally, visual information may provide semantic information to resolve structural ambiguities. Word quintuples such as *(a) hotel caught fire during (a) storm* were extracted from English Wikipedia and the attachment locations were automatically labeled either as 'n' for low attachment, where the prepositional phrase adjoins the direct object, or 'v' for high attachment, where the prepositional phrase adjoins the main verb (Nakashole and Mitchell, 2015). 168 test items are selected by human annotators for evaluation, within which 119 are sentences with high attached PPs and 49 are with low attached PPs. For evaluation of PP attachment with induced trees, one test item is labeled correct when the induced tree puts the main verb and the direct object into one constituent and it is labeled as 'v'. For example, if the induced tree has *caught fire* as a constituent, it counts as correct for the above example with high attachment. Low attachment trees must have a constituent with the direct object and the prepositional phrase. For example, for the sentence *(a) guide gives talks about animals*, the induced tree must have *talks about animals*. Average accuracies for all sentences as well as for sentences with high attachment or low attachment with induced grammars are shown in Figure 5. Results show that the models trained with visual information on both datasets show significantly higher performance on the PP attachment task in most of the categories, except for the low attachment category with Multi30k models where the performance from both models is not significantly different. This is in contrast to the

Models	MSCOCO				Multi30k					
	En	Ko	Pl	Zh	De	En	Fr	Ko	Pl	*Zh*
NoImagePCFG	**46.2**	21.7	**45.8**	46.0	**52.8**	**49.9**	42.2	22.8	38.9	*51.6*
ImagePCFG	41.2	**26.4**	40.2	**48.1**	51.3	39.9	**42.6**	**33.2**	**39.7**	***53.2***

Table 3: Average labeled Recall-Homogeneity of the NoImagePCFG and ImagePCFG models on the MSCOCO and Multi30k caption datasets with random words replaced by the UNK symbol. Standard deviations across the datasets are similar to what is reported in Table 1 and 2. Chinese Multi30k is the one on which the NoImagePCFG model outperforms the ImagePCFG model on the normal test set but not on the UNK test set.

higher performance of the NoImagePCFG models on unlabeled F1 and labeled RH than that of the ImagePCFG models on English from both caption datasets. Results indicate that induction models use visual information for weighting competing latent syntactic trees for a sentence, which is consistent with the third hypothesized advantage of visual information for induction. This also indicates that the reason that the overall parsing performance of ImagePCFG on English is lower than NoImagePCFG lies within other syntactic structures, which is left for future work.

8 Conclusion

This work proposed several novel neural network-based models of grammar induction which take into account visual information in induction. These models achieve state-of-the-art results on multilingual induction datasets without any help from linguistic knowledge or pretrained image encoders. Further analyses isolated three hypothesized benefits of visual information: it helps categorize noun phrases, represent unknown words and resolve syntactic ambiguities.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. Computations for this project were partly run on the Ohio Supercomputer Center (1987). This work was supported by the Presidential Fellowship from the Ohio State University. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. This work was also supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & Prepositional attachment resolution in captioned scenes. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1493–1503.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Andrew Drozdov, Patrick Verga, Yi-Pei Chen, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised labeled parsing with deep inside-outside recursive autoencoders. 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 1507–1512, Hong Kong, China. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language development: Vol. 2. Language, thought, and culture*, 2(1):301–334.

Dedre Gentner. 2006. Why Verbs Are Hard to Learn. In K. Hirsh-Pasek and R. Golinkoff, editors, *Action Meets Word: How Children Learn Verbs*, pages 544–564. Oxford University Press.

Lila Gleitman. 1990. The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55.

Ajda Gokcen, Ethan Hill, and Michael White. 2018. Madly ambiguous: A game for learning about structural ambiguity and why it's hard for computers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 51–55, New Orleans, Louisiana. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 2, pages 725–731.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018a. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2721–2731, Brussels, Belgium. Association for Computational Linguistics.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018b. Unsupervised grammar induction with depth-bounded PCFG. *Transactions of the Association for Computational Linguistics*, 6:211–224.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019. Unsupervised learning of PCFGs with normalizing flow. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452, Florence, Italy. Association for Computational Linguistics.

Lifeng Jin and William Schuler. 2019. Variance of average surprisal: A better predictor for quality of grammar from unsupervised PCFG induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2463, Florence, Italy. Association for Computational Linguistics.

Lifeng Jin and William Schuler. 2020. The Importance of Category Labels in Grammar Induction with Child-directed Utterances. In *Proceedings of 16th International Conference on Parsing Technologies*, Seattle, USA. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov chain Monte Carlo. *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–146.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019a. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence, Italy. Association for Computational Linguistics.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019b. Unsupervised recurrent neural network grammars. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1105–1117, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3686–3693.

Ndapandula Nakashole and Tom M Mitchell. 2015. A knowledge-intensive model for prepositional phrase attachment. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 1, pages 365–375.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan Mcdonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference*.

Hiroshi Noji and Mark Johnson. 2016. Using Left-corner Parsing to Encode Universal Structural Constraints in Grammar Induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 33–43.

The Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. \url{http://osc.edu/ark:/19495/f5s1ph73}.

Steven Pinker and B MacWhinney. 1987. The bootstrapping problem in language acquisition. *Mechanisms of language acquisition*, pages 399–441.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the 4th International Conference on Learning Representations*. International Conference on Learning Representations, ICLR.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Yoav Seginer. 2007. Fast Unsupervised Incremental Parsing. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

Cory Shain, William Bryce, Lifeng Jin, Victoria Krakovna, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2016. Memory-bounded left-corner unsupervised grammar induction on child-directed input. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 964–975, Osaka, Japan. The COLING 2016 Organizing Committee.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In *ICLR*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks. In *ICLR*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.

Michael Tomasello. 2003. *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press, Cambridge, MA, US.

Sandra Waxman, Xiaolan Fu, Sudha Arunachalam, Erin Leddon, Kathleen Geraghty, Hyun-Joo Song, Child Dev, and Perspect Author. 2013. Are Nouns Learned Before Verbs? Infants Provide Insight into a Longstanding Debate NIH Public Access Author Manuscript. *Child Dev Perspect*, 7(3).

A Details of datasets

The MSCOCO caption dataset used in Shi et al. (2019) contains 413,915 sentences in the training set, and 5000 sentences in the development and test sets respectively.[12] Every image is accompanied by 5 captions,and there are 82,783 images in total in the training set. The image embeddings of size 2048 used in Shi et al. (2019) are encoded by an image classifier with ResNet128 architecture trained with on the ImageNet classification task (Deng et al., 2009).

The Multi30k caption dataset contains 29,000 sentences in the training set, and 1,014 sentences in the development and 1,000 in the test set in four different languages, all of which except Czech are used in this work thanks to the availability of high accuracy constituency parsers in these languages.[13] There are as many images as there are captions in the training set. The image embeddings of size 2048 provided with the dataset are encoded by an image classifier with ResNet50 architecture also trained with on the ImageNet classification task.

For data preprocessing, following Shi et al. (2019), the size of the vocabulary is limited to 10,000 for all languages and datasets. All raw images are resized to $3 \times 64 \times 64$ and normalized with means $[0.485, 0.456, 0.406]$ and standard deviations $[0.229, 0.224, 0.225]$, calculated from images in ImageNet.

B Hyperparameters

The hyperparameters used in all proposed models are tuned with the MSCOCO English development set. For the grammar induction model, the size of word and syntactic category embeddings, as well as the size of hidden intermediary representations is 64. The size of the image embedding in the ImagePCFG system is also 64. All out-of-vocabulary words are replaced by the UNK symbol. Sentences with more than 40 words in the training set are trimmed down to 40 words. For the projector model, five different convolutional kernels, from (1,64) to (5,64), are used with 128 output channels. The trainable image encoder employs a ResNet18 architecture,[14] and the decoder employs the decoder architecture in the DCGAN model.[15]

A batch size of 2 is used in training. Adam is used as the optimizer, with the initial learning rate at 5×10^{-4}. The loss on the validation set is checked every 20000 batches, and training is stopped when the validation loss has not been lowered for 10 checkpoints. The model with the lowest validation loss is used as the candidate model for test evaluation, where best parses are generated with the Viterbi algorithm on an inside chart.

C Development

Table 4 and 5 report unlabeled F1 and labeled RH results on the development sets in the multilingual caption datasets. Results show that development and test results are very similar, indicating that the general characteristics of the two sets are very close.

[12] The data set can be found at https://github.com/ExplorerFreda/VGNSL along with image embeddings encoded by pretrained image encoders.

[13] The data set can be found at https://github.com/multi30k/dataset along with image embeddings encoded by pretrained image encoders.

[14] https://pytorch.org/docs/stable/_modules/torchvision/models/resnet.html#resnet18

[15] https://github.com/pytorch/examples/blob/master/dcgan/main.py

Models	English		Korean		Polish		Chinese	
	F1	RH	F1	RH	F1	RH	F1	RH
NoImagePCFG	60.3±8.2	46.4±11.0	38.6±8.7	22.6±6.9	59.5±3.8	47.5±3.9		
ImagePrePCFG	55.7±7.5	39.6±5.4	39.5±4.2	24.1±3.4	61.2±1.6	50.1±3.3		
ImagePCFG	55.4±2.7	43.2±1.8	45.1±2.3	27.5±2.6	54.3±8.3	41.6±7.9		

Table 4: Averages and standard deviations of labeled Recall-Homogeneity and unlabeled F1 scores of various unsupervised grammar inducers on the MSCOCO caption development datasets.

Models	German		English		French	
	F1	RH	F1	RH	F1	RH
NoImagePCFG	47.2±5.7	53.6±5.7	59.1±8.1	52.2±8.5	43.8±4.9	43.2±5.2
ImagePrePCFG	44.8±7.9	50.0±8.3	46.7±7.3	40.7±7.5	42.3±10.3	42.8±10.5
ImagePCFG	45.6±5.2	50.6±8.5	47.7±5.4	40.9±5.2	43.1±5.1	43.9±5.5
Models	Korean		Polish		Chinese	
	F1	RH	F1	RH	F1	RH
NoImagePCFG	30.6±5.7	22.2±3.0	49.4±4.9	40.0±5.3	59.7±3.3	53.6±4.7
ImagePrePCFG	27.0±4.8	19.2±3.6	48.5±3.1	38.5±3.1	55.5±9.3	48.3±10.4
ImagePCFG	45.1±1.1	33.4±1.9	49.5±7.6	40.8±6.3	58.3±3.2	52.1±4.3

Table 5: Averages and standard deviations of labeled Recall-Homogeneity and unlabeled F1 scores of various unsupervised grammar inducers on the Multi30k caption development datasets.