What Can We Learn from the Structures Found in Visual and Language Data and their Correlations?

Opening the Discussion on Joint Visual and Linguistic Structures

Victor Milewski (PhD), Dr. Maria Mihaela Trusca (Postdoctoral researcher) and Prof. Dr. Marie-Francine Moens (Professor)

KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium

Abstract

Structure is omnipresent in the physical world and in languages that humans use to communicate about the world. Structure refers to the arrangement and organization of elements. For language, these structures have been well defined in the form of grammars, while for visual data it is still uncertain how objects are structured in the physical world. In this work, we describe and discuss the known linguistic grammar systems and correlate them to what has been discovered about visual structures. This could lead to better representations of content that capture a joint structure and their use in artificial intelligence systems. A preliminary analysis is performed to show the correlations between the spatial distances between objects in the image and the grammatical and semantic structures found in the image descriptions created by humans. These show that syntax trees, that is, constituency and dependency trees, are correlated to the spatial structure of the image. Scene graphs and scene trees that during their creation are informed by the image, have the best correlation. To further inspire research into representations that integrate structure in content representations of the physical world, we discuss different areas where the processing of visual data is researched. These include 1) psychology studies where brain responses for visual semantic inconsistencies are tested, 2) developed scene graph annotations forming a descriptive language, and 3) studies evaluating structure in foundation models. Finally, we state some open questions about structures found in visual data to direct future studies on content representations of the physical world and their language descriptions and on the understanding of them by the machine.

Keywords

Scene grammar, Text grammar, Scene graphs, Visual understanding, Structural analysis, Representational Similarity Analysis,

1. Introduction

The physical world that we humans perceive is extremely varied and is composed of many different types of objects with their own perceptual properties. Objects can be animated and function as actors (e.g., humans, animals) or inanimate (e.g., books, tables) that require interaction with an actor to change position. Many of the inanimate objects are placed and organized in the world by actors. Objects have interactions between them. They might refer to their respective spatial positions, but it might also be an action applied that interacts with the object. Some types of objects could only have certain types of interactions and can never show any relation where they influence each other. Take, for example, the sky and a cloud, the cloud can move through the sky, by it doesn't change the sky. This is different for actors, they can have both spatial relations, but they can also interact with some other objects and influence their properties.

When observing the physical world, humans easily learn to identify and understand the relations between different objects. Humans often do not reason about many of these interactions as they are extremely common. This relates to a commonsense understanding of the world, which directs certain expectations of object relationships and behavior. When something goes against an expectation, it can draw the attention of humans, and the unexpected relation is somehow highlighted in the processing of the brain [1, 2, 3, 4, 5]. These examples highlight an important feature of human processing of the physical world. Humans tend to learn and assign specific structures to the world around them [6].

Humans use language to communicate about the physical world. Natural language is here the most prominent and common form. Natural language typically has a structure that is guided through the use of grammar rules. Many of the grammar rules are universal and found in many languages and cultures [7]. Humans also use a formal language (e.g., in the form of object entities, their attributes, and relationships) to describe the physical world which might make it easier to interface with

MRC 2023 – The Fourteenth International Workshop Modelling and Representing Context. Held at ECAI 2023, 30.09.-5.10.2023, Kraków, Poland.

✉ victor.milewski@kuleuven.be (V. Milewski); mariamihaela.trusca@kuleuven.be (M. M. Trusca); sien.moens@kuleuven.be (M. Moens)
🌐 https://calculus-project.eu/team/victor.html (V. Milewski); https://calculus-project.eu/team/maria.html (M. M. Trusca); https://people.cs.kuleuven.be/~sien.moens/ (M. Moens)
0000-0003-3854-6386 (V. Milewski); 0000-0002-9204-0389 (M. M. Trusca); 0000-0002-3732-9323 (M. Moens)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

machines and programming code. Language, whether it is natural or formal, imposes a structural graphical representation on the content it conveys.

Artificial Intelligence research targets human-like understanding of images and language as well as the generation of images and language that truthfully simulate the physical world and language uttered by humans, respectively. AI research has recently witnessed large advancements concerning the above tasks due to the development of foundation models (e.g., language models, vision models, visio-linguistic models [8, 9, 10]) and neural transformer architectures [11] that make it possible to detect relationships in visual and textual data by means of sophisticated attention mechanisms. Notwithstanding, there remains a need to more explicitly represent structure that naturally is present in visual and language data, as this could lead to representations (e.g., obtained with deep neural networks) that improve the compositionality of the representations and the reasoning with these by the machine. Compositionality is linked to human's ability to produce and interpret novel utterances in language or construct novel compositions of objects.

All this leads to a number of research questions.

- How does the natural structure found in visual data (e.g., spatial, temporal, causal structure) aid language understanding and generation by the machine?
- How does the structure of language data help processing visual data?
- Is there a joint structure that we can capture in the content representations that might aid both the processing of language and visual data?
- Can we induce that structure automatically from the data?

In this paper, we do not yet formulate final answers to these challenging questions but perform experiments that are a step towards their answers and that could form the basis for a scientific discussion on the topic.

The remainder of the paper is structured as follows. We first discuss the formalisms selected to structure language that describes the content of an image, constituency trees, dependency trees, scene trees, and scene graphs (Section 2). The next sections discuss the structure of visual data and how we can detect a joint structure shared by the visual and language data (Sections 3 and 4). Then, we discuss the experimental set-up of our preliminary experiments and report the results and their discussion (Section 5). Finally, we discuss and analyze the differences and commonalities between these structures leading to important research questions for future work (Section 6).

2. The Structure of Language

In this paper we consider the grammatical structures found in natural language used to describe images, as well as the semantic graphical structure drafted in a controlled language used to describe images. More specifically, we consider four structural formalisms: constituency trees, dependency trees, scene trees, and scene graphs. They all build a structured representation of language content, in most cases restricted to one sentence. Constituency trees and dependency trees are solely based on the natural language sentence. A scene tree uses both language and visual input to create the structured representation of the natural language sentence reduced to the objects present in the image. The scene graph uses a controlled language vocabulary of object labels and relationships to structurally describe the content of an image.

2.1. Constituency Trees

Constituency trees, also known as phrase structure trees or syntactic trees, are graphical representations used in linguistics to analyze the structure of sentences. An example of a constituency tree is shown in Figure 1 on the right. The constituency tree shows how words and phrases in a sentence are grouped together into larger constituents or phrases. As a result, the sentence is represented as a hierarchical structure with multiple levels. At the highest level, there is a single node called the root, which represents the entire sentence. The root node branches out into smaller nodes, each representing a constituent or a phrase within the sentence. The nodes can be of different types, such as noun phrases (NP), verb phrases (VP), and prepositional phrases (PP), forming part-of-speech tokens. The leaf nodes of the tree represent the individual words of the sentence. Formally, a constituency tree is a simple, undirected, connected, acyclic graph G. The structure $G = (V, E)$ consists of a set of vertices V, and a set of pairs of vertices E, which we refer to as arcs or edges. A tree with n nodes has $n - 1$ graph edges that form the tree structure.

A constituency tree can be described by a a Context-Free Grammar (CFG) [13], which is a mathematical system for modeling constituent structures of natural languages. As described above the constituent is a group or unit of words that can behave as a single structure, such as a noun phrase, and grammars based on them were already theorized by Wundt [14]. The CFG has a set of rules that describe how to combine words and phrases and how they are ordered together.

Constituency trees are used in varying tasks and can easily be automatically derived from the text. Some of the more traditional parsers use rule-based methods, while nowadays mainly automatically trained deep learning models are used [15, 16, 17, 18]. These self-trained meth-

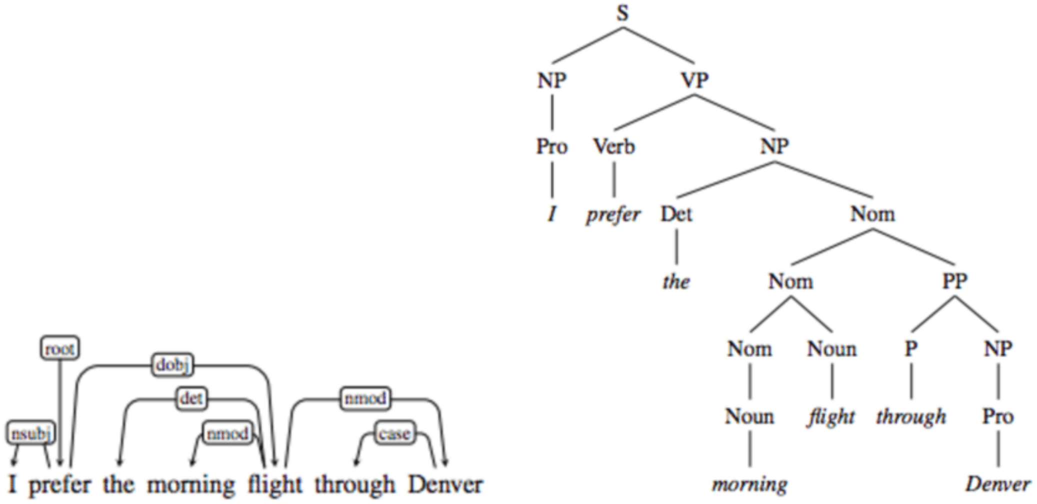

Figure 1: A comparison for the parse trees made through a dependency grammar on the left and a constituency grammar on the right. Parse trees created by Jurafsky and Martin [12].

ods make use of conditional random fields (CRFs), Recurrent Neural Networks (RNNs), transformer architectures, and other attention-based networks with an F1 score recognition of the constituents of up to 96%. Note that the parsers are not perfect yet, due to ambiguity in the language and grammar it can lead to multiple interpretations of a single sentence, thus resulting in different parse trees. In the experiments below, where we report on parsing simple image captions, we select the most probable parse tree.

2.2. Dependency Trees

A dependency tree is another graphical representation of the structure of a sentence used in linguistics to analyze the syntactic relationships between its words. Unlike constituency trees that focus on grouping words into phrases, dependency trees emphasize the grammatical relationships between individual words. An example is shown in Figure 1 on the left. The relations between a word and its dependent are not limited to the word order in the sentence and they can be further apart in the sentence, creating a flatter structure. In a dependency tree, each word in the sentence is represented as a node, and the relationships between words are represented as directed arcs or edges between two nodes. The nodes include the word itself, along with additional information such as its part of speech (POS) tag. Relationships are usually labeled with their grammatical role that describes the nature of the dependency, for instance, subject, object, modifier, adverbial, and others.

Formally, a dependency structure is a directed graph. The structure $G = (V, E)$ consists of a set of vertices V, and a set of ordered pairs of vertices E, which we refer to as arcs or edges. The set of vertices V corresponds to the set of words in a given sentence. The set of edges E captures the head-dependent and grammatical role relationships between the elements in V. A dependency tree satisfies the following constraints. There is a single designated root node that has no incoming arc; with the exception of the root node, each vertex has exactly one incoming arc; here is a unique path from the root node to each vertex in V.

A dependency tree can be described by a dependency grammar [19]. A large effort has been made to create a standard towards a universal grammar regarding dependencies [20, 21]. This resource contains treebanks for many languages, with the idea that dependents are similar between multiple languages. This allows for the transfer of knowledge between languages and can aid in translation models.

Automatic methods have been developed for generating the dependency parse of a sentence as well, i.e. the shift-reduce algorithm [22]. More advanced algorithms are designed using deep learning methods such as bidirectional Long Short-Term Memory Networks (BiLSTM) graph-based parsers [23], graph neural networks [24], and attention-based models [25] achieving a labeled attachment score for dependencies of up to 96%. Interestingly, in recent exploratory studies that investigate the knowledge of foundational models, it is found that they automatically learn the language grammar without being trained for it [26]. With the use of structural probes, they found that the trained embeddings contain

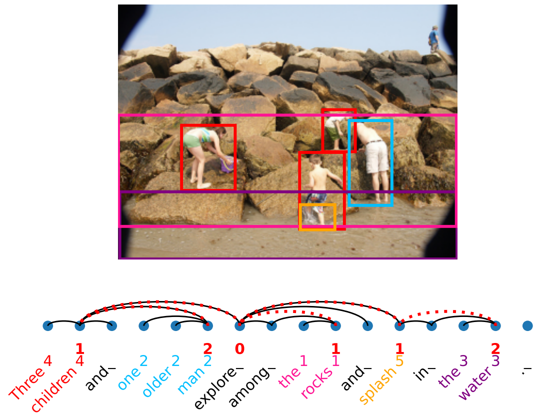

Figure 2: An image shown with its caption and the parsed dependency tree. In the red-dotted lines, the constructed scene tree [27] indicates the dependent relations between objects in the image.

strong knowledge about dependent relations and they use specific layers for storing this knowledge.

2.3. Scene Trees

Based on the discovery by Hewitt and Manning [26], that the grammatical structure of language is automatically detected in the BERT model [8], Milewski et al. [27] hypothesized that multimodal BERT models could do the same for the visual data. As described in Section 3, there is evidence that humans process grammatical structures for different modalities, such as language and visual scenes in a similar way. In order to evaluate the presence of structure encoded in the embeddings for objects of a scene, the scene tree was proposed [27]. The scene tree follows the dependency parse tree, but is truncated to just include the objects and elements also present in the visual data. An example of this is shown in Figure 2. Formally, a scene tree is a directed graph comparable to a dependency tree. The structure $G = (V, E)$ consists of a set of vertices V, and a set of ordered pairs of vertices E called edges. However, the set of vertices V is restricted to the set of words in a given sentence that are aligned with an object in the image. The set of edges E only captures the head-dependent relationships between the elements in V.

The scene tree is constructed by starting from an empty scene tree, with at its root the full image. Next, the sentence is traversed along the edges of the dependency tree. Whenever a head noun is discovered that aligns with a region in the image, it is added to the scene tree by attaching it to the last added element (in the first step the root). By traversing along the dependencies, a perfect alignment between both trees is ensured.

Milewski et al. [27] found that the embeddings obtained with the multimodal BERT model do not well capture structural knowledge concluding that the transfer between structures of language and objects is currently not occurring in visio-linguistic foundation models, and the training objective is not properly designed to achieve this. These results are confirmed by the work of Hendricks and Nematzadeh [28], where *SVO-probes* are designed to test the capabilities of subject-verb-object triplet understanding in models. Here they found that especially the understanding of verbs (or relationships) is difficult for such models. This finding is confirmed by Bugliarello et al. [29] who highlight the limitations of the foundation models in fine-grained tasks, where a precise understanding of structures in both modalities is needed, and implemented a next attempt to train models capable of solving them. Examples of such tasks include verb understanding [28], word order [30], spatial relations [31], and other linguistic phenomena [32]. These studies reveal that there is still a big opportunity to further investigate how compositionality works in visual data and how it correlates and interacts with linguistic grammar.

Nevertheless, the scene tree [27] forms a valid formalism for representing the caption of an image, hence its use in our experiments.

2.4. Scene Graphs

Our languages are important for describing the world, we assign names and labels to objects and when we talk about the environment around us, we need to use language. When designing visual graphs that describe the world, it must use text to assign labels to the vertices and nodes. This indicates that the graph on its own is described by a unique language that we can learn and structure with rules. A common controlled and structured form of language is represented by knowledge graphs that have the quality of capturing explicit knowledge about the physical world. To assure the mapping of the knowledge graph to the real world, the nodes of the graph become entities of interest while the edges represent the semantic relations between these entities.

One of the most commonly used visual graph structures are the Visual Genome (VG) scene graphs [33]. Formally, given a set of object classes C, a set of attribute types A, and a set of relationship types R (in our case it is a binary indicator for the presence of a relationship between two nodes), scene graph G is a tuple $G = (V, E)$ where vertices $V = o_1, ..., o_n$ is a set of object labels and $E \subseteq V \times R \times V$ is a set of edges. In a scene graph, each object o_i is defined as follows:

$$\forall_{o_i \in O}, o_i = (c_i, A_i), \text{with } c_i \in C \text{ and } A_i \subseteq A$$

Where A_i are the attributes of the object o_i.

Such scene graphs are constructed based on an image, however, the graph by itself is not related to an image. It describes a scene that can be depicted by an image (or more). To make it a scene graph, it must be grounded to the images bounding boxes *B*.

The object instances in the scene graph can be a person, a place, a thing, or parts of other objects. Attributes are used to describe the state of the current object; these may include its shape, color, and pose. Relations are used to describe the connections between pairs of objects, such as actions, and positions.

VG scene graphs were designed based on two arguments, 1) Explaining relations is *cognitive* in nature, and 2) datasets with simple object box annotations (MS-COCO or VQA) do not allow distinguishing two images when both have the same objects. To accommodate a better understanding of images two of the key elements are 1) the grounding of visual concepts to language and 2) the specified formalized representations of the components of an image.

With these goals and reasons, it is clear that they required the graphs to be as densely annotated as possible, otherwise, two images might not be distinguishable through the graph.

Here we will describe the process taken to collect the data for the VG scene graphs. The annotations were made in six steps and all annotations were verified and canonicalized (to WordNet synsets [34]) by multiple workers.

Data collection process The data collection process started with workers annotating images with natural language descriptions. They are short descriptions of the content of a region instead of full image captions. Multiple workers create three regions per image each, collecting at least 50 descriptions per image. To ensure a varied set of descriptions, it was enforced that they are not too similar to others of the same image and that they differ enough from the top hundred occurring descriptions across all images in the dataset. This ensures that descriptions do not repeat often across many images and it can help to make more descriptive and specific descriptions.

Figure 3 is an example of an image and all its region annotations. These annotations clearly describe many of the small details in the image.

For the object annotations, the workers receive an image and one region description (from the previous step). The annotator has to discover all the mentioned objects in the text and localize them in the image by drawing a bounding box around them. If the description in the image matches an existing box, they can join the annotations, helping with the combining of sub-graphs of the image in a later step.

With all the objects annotated, the next step is to annotate the attributes and relationships. Given the image, a description, and annotated boxes for the description, the worker has to assign mentioned attributes (such as color and state) to the objects. Furthermore, described relationships between two mentioned objects must be annotated. For example in the first description on the right side in Figure 3, the worker would assign the attribute *black* to the ”shirt” and draw the relation *on* between ”shirt” and ”man”. Note that these relationships are directed.

Region and scene graph creation With all the annotations collected, they are merged into region graphs per description. For each bounding box, every attribute, and every relationship, nodes are created of a unique type each. For each attribute node, an edge is drawn from the object node that it belongs to. For each relationship node two edges are drawn, one from the subject node to itself and one from itself to the object of the relation.

With all the region graphs constructed, they are combined into one scene graph for the entire image (see Figure 4 for an example of the process). Graphs are merged based on the joined object annotations made earlier, and by computing if boxes have an intersection over union higher than 0.9.

Usage and automatic generation of scene graphs An increasing amount of research moves towards the automatic generation of these scene graphs [35, 36, 37, 38, 39]. Interestingly, some of these works aim to learn to generate solely on language supervision [40]. They create a multimodal transformer model [11], that receives triplet words `<subject, predicate, object>` from a caption and a set of predicted bounding boxes. A word is masked for the triplet and the model is trained to predict all triplet words. The triplets from the caption are collected through a language parser and aligned to the predicted bounding boxes based on the labels. The used graph parser [41], which is based on the parser by Schuster et al. [42]. These scene graph parsers are again based on dependency parsers, as described in Section 2. Instead of labeling SG nodes, they label the edges with attribute, subject, or object labels, which makes the SG more related to the dependency tree. they align the graph with the tree by aligning object labels with words based on synonyms.

The scene graphs are also used for improving the training of visio-linguistic models [29]. By creating very simple captions from triplets in the graph, a dataset is formed that can be used for masked-language modeling. this improved fine-grained understanding of the models and they achieved state-of-the-art spatial reasoning abilities, showing the benefit of strong structured representations of visual data.

The above four graph structures will be used in our

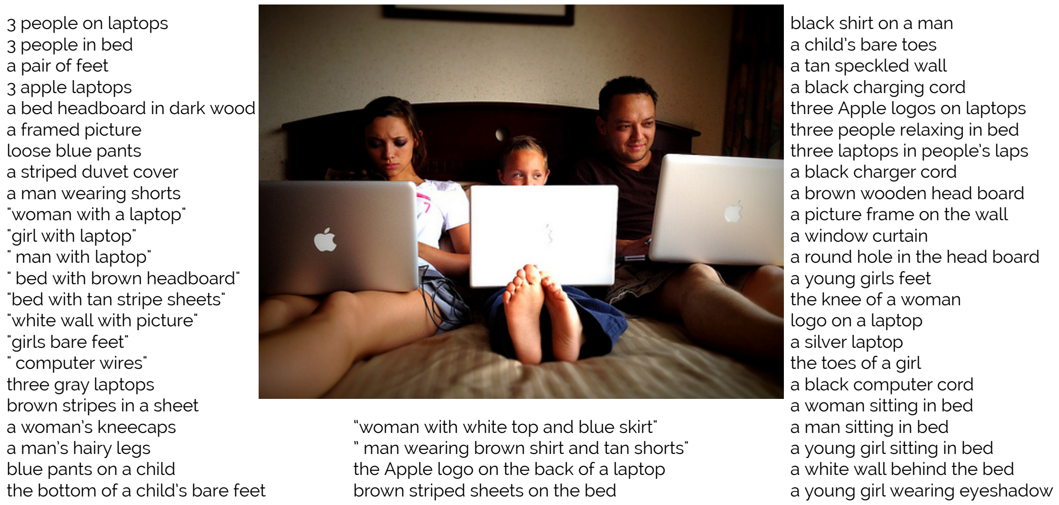

Figure 3: Example image with all its region descriptions used to create the sub-graph parts of the scene graph.

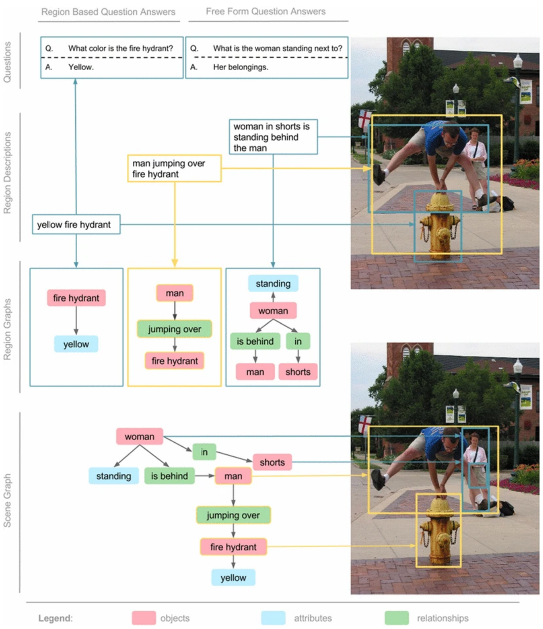

Figure 4: Example of the process for combining the region graphs into the full scene graph of the image. Image taken from [33].[1]

experiments below as structural representations of the language descriptions of an image. As an overarching principle, the structure will be captured by the distance between a pair of nodes when there is a path that connects the nodes in the graph structure. This distance is computed by the path length of the shortest path in the graph. When computing this path length we do not take into account the direction of the edges, if present, nor the grammatical or semantic labels of the relationships between nodes. We leave it to future work to use this information when finding correlations between language and visual structures.

3. The Structure of Visual Data

One important aspect of understanding the structure of the physical world is how humans perceive and process it. Different works have been conducted to understand how humans process visual data and how they design the physical world around them. These works often investigate scene-based visual search. Dependent on what they are searching for, humans might look at specific objects, or use global information about the scene [43]. When looking, for example, for bread in a kitchen scene, people have semantic and episodic knowledge. The former guides the user in their understanding of the scene and environment and what likely places are for bread, while the latter suggests a familiarity with the scene, i.e. someone finds the bread quicker in their own kitchen. This work by Wolfe et al. [43] already hints that humans create a quick semantic interpretation of a scene, after which they can do more efficient processing.

When further exploring the human understanding of visual scenes, a study was done on memorizing objects and their locations in a natural scene [44]. A comparison was made where some subjects were asked to memorize the scene, while others were tasked with searching objects. The latter group had a much higher recall. This

[1] Link to CC-BY license.

benefit was however lost when a non-descriptive background was used with randomly placed objects on it. Draschkow et al. [44] conclude that scene semantics help to produce a representation that supports human memory. Le-Hoa Võ and Wolfe [45] describe this knowledge of scenes linking it to episodic and semantic memory. Humans use these different types of memory, where the former can be described as having seen a room before, and the latter as being familiar with common patterns and positions in rooms. These common patterns and positions can be described as a *scene grammar*. For example, it is to be expected that different types of cutlery are close to each other in a scene and that the knife is usually on the right of the fork.

The above has led to approaches that impose a grammatical structure on a visual scene. This has been on different scales, for example, only on objects, on specific structures such as roads, or very general on entire scenes. Here we discuss some of these structures and analyze their design goals and applicability. One such grammar was created to analyze the design of general objects, like chairs and tables [46]. The grammar is trained such that it parses objects into parts. It is trained jointly with a very descriptive caption using a contrastive loss between the image and the sentence parse tree. Another grammar was constructed to form road networks [47, 48]. Here a grammar is trained from annotated image data of roads. Based on this, the rules are learned for road layouts. For instance, given a city road, there must be sidewalks, maybe a bicycle lane, and the roads are a bit more narrow. Using these grammars, they can create better 3D generations of roads for experimenting in simulations.

An alternative approach to scene grammar studies is to ask participants to place objects in a room and probe their memory afterward [49]. By asking participants to either place objects according to grammar or randomly, and afterward do a surprise recall test where they had to reconstruct it, they found that the processing time in the random setting was much higher. Furthermore, they noticed that during building the participants tend to grab the larger global objects early on and the smaller local objects later. This hints a the structure of the grammar is based on constituencies. An example interpretation of a scene grammar is shown in Figure 5. First, the larger objects of a room are placed, such as the sink and the bathtub, later the smaller items are placed around it in phrases such as the *sink phrase*, which holds items like toothpaste and a toothbrush. Thus, the global objects are used as "anchors" that can guide visual search [49].

The above works show that the structure or grammar that can be recovered from images is primarily spatial. For this reason in our preliminary experiment described below, we rely on distances between objects in the 2D visual scene. For building a structural representation of the 2D visual scene, we consider its objects and their spatial positions in the scene. Objects are represented by their bounding boxes where we consider the 2D coordinates of the center of the bounding boxes. We assume that the structure of a visual scene is determined by objects that interact and consider the distance between objects in the scene as a measure of this interaction. We do not normalize the size of the object bounding boxes, as in the experiments below we will compare the structure of a visual scene with a selected type of language structure that describes that image, and do not compare object boxes across images. As a distance metric, we use the Euclidean distance (in image pixels) between the center of two bounding boxes.

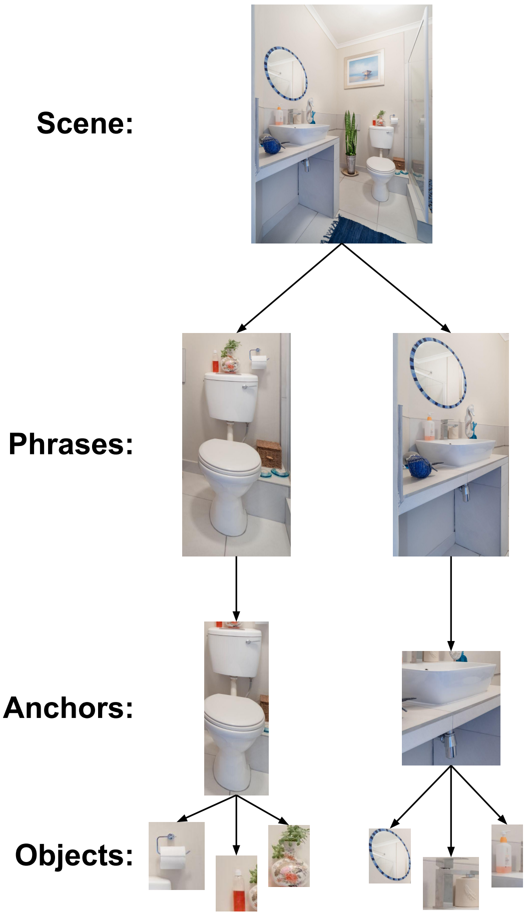

Figure 5: Example of a scene grammar schematic hierarchy of a bathroom scene. It shows the early placement of large ”anchor” objects and the grouping of smaller objects around them as ”phrases”. Figure recreated following [6].

4. Finding Correlations between the Structure of Visual and Language Data

There is no research or proof yet on finding correlations between the structures found in visual and language data. However, when looking at different Electroencephalogram (EEG) studies that discover brain responses, similar brain responses were found when subjects saw semantic inconsistencies in visual or linguistic data [1, 2, 3, 4]. Similarly, a correlation between the brain responses can be found when examining syntactic inconsistencies in both language and visual data [3, 5]. Noteworthy is that in the study by Cohn et al. [5], they made use of constituency structures describing the narrative of comics. This shows that grammatical systems could exist in other modalities, such as the visual narrative, and that inconsistencies in the syntax are processed similarly to text. Other modalities where grammars are proposed are in drawing [50], music [51], and in personal relations [52]. Grammatical structures are thus strongly present in human understanding and processing in many different paradigms.

While there are indications that the brain responses can be similar regarding semantic and syntactical structures for language and other modalities, it is not evident yet that the processing is equal. However, Võ et al. [6] states this is most likely a continuum that requires further investigation.

In the preliminary experiments below we study the correlations found between the detected structure of the visual scene and the detected structure of its language description. For the latter, we consider its constituency tree, its dependency tree, its scene tree, and its scene graph.

5. Preliminary Experiments

5.1. Data Set

Our data set is composed of 483 captions assigned to 145 images selected from the Flickr30K dataset [53][2] with the Entities extension [54][3]. The image-caption pairs were selected because of the overlap with the images and corresponding scene graphs of Visual Genome [33][4]. With the Flickr30K-entities we have a direct alignment between nouns in the caption and object bounding boxes in the image, which is needed when we compare image structures with dependency trees, constituency trees, and scene trees obtained from the image captions. The captions were parsed into constituent and dependency trees with the Spacy parser[5] and the Berkley neural constituency parser [55][6]. For obtaining the scene trees we relied on the Multi-modal Probes code base [27][7].

5.2. Correlations between the Language Structure and the Visual Structure

The most interesting experiments regard the cross-modal correlations, where we compare the structure found in the images and the structure in the language descriptions. Here we compare two graph structures each composed of nodes and edges, and the distances between two nodes in the graph structure as defined above. This allows us, for instance, to explore how well the distance between objects in the physical world is captured by language. The distances computed between objects in an image are compared with the distances of objects in the language graphs that describe the images. In the cases where dependency trees, constituency trees, and scene trees are used as structural language descriptions, the shortest path between head nouns inside the caption is computed. In the case of scene graphs, we compute the shortest path between labeled nodes of the scene graph.

When computing correlations, we mainly rely on representational similarity analysis (RSA) [56]. This analysis technique allows to quantitatively compare measurements of different modalities. The first step of the RSA is to derive the representational dissimilarity matrices (RDMs), which characterize the information carried by a given representation (in our case the RDMs contain the distances between objects computed in the respective modality). Then, for each RDM, we select and vectorize its upper triangular matrix (excluding the diagonal) to then calculate Spearman's rank correlation ρ between the vectorized RDMs. Furthermore, we report Spearman's rank correlation directly over the distances as well.

We will express the RSA and Spearman's rank correlation by analyzing their distribution of results for all the graph-image pairs. We report the scores at the three quantiles, so 25% of the pairs have a score below the one at Q1, 50% of the pairs below the score at Q2 (the median), and 75% of the pairs below the score at Q3.

5.2.1. Correlations between the Visual Structure of the Image and the Constituency Parse of the Caption

We create the constituency tree of each image caption and compute the distance between the head nouns connected through the constituency tree. The distribution

[2] https://shannon.cs.illinois.edu/DenotationGraph/
[3] https://bryanplummer.com/Flickr30kEntities/
[4] https://paperswithcode.com/dataset/visual-genome
[5] using the en_core_web_md model https://spacy.io/
[6] https://spacy.io/universe/project/self-attentive-parser
[7] https://github.com/VSJMilewski/multimodal-probes

Table 1

The quantiles of the Spearman rank correlation and the RSA metric computed considering the distance between objects in the image and the distance between their corresponding head nouns in the image caption, according to the constituency tree.

Spearman Correlation			RSA		
Q1.	Q2	Q3	Q1.	Q2	Q3
0.49	0.76	0.91	-0.03	0.55	0.81

Table 2

The quantiles of the Spearman rank correlation and the RSA metric computed considering the distance between objects in the image and the distance between their corresponding head nouns in the image caption, according to the dependency tree.

Spearman Correlation			RSA		
Q1.	Q2	Q3	Q1.	Q2	Q3
0.51	0.75	0.9	-0.03	0.53	0.81

of correlation scores between distances for objects in the image and path distances in the tree are shown in Table 1. The correlation between the image and the constituency tree is in general positive, which confirms our assumption that the increased distance between the head nouns in the constituency tree represents an increased physical distance between the objects of the real world. However, the strength of the correlation is modest, even if the third quantile is quite high for both RSA and Spearman rank correlation.

5.2.2. Correlations between the Visual Structure of the Image and the Dependency Parse of the Caption

We create the dependency tree of the image caption and compute the distances between the connected head nouns in the tree. We compare these with the distances between the objects that correspond to the head nouns in the image. The distribution of correlation scores is shown in Table 2. The correlations of the distances are very similar to the correlations with the constituency trees from the previous subsection. The correlation is positive meaning that increasing the distance between the head nouns of the dependency tree is represented in the real world by an increased distance between objects. Again the correlations are moderate indicated by the medians at 0.5 and 0.75 for the RSA and Spearman rank correlation. However, the moderate correlation can be expected especially considering that sometimes we can indicate the location through language by connecting an object with a reference place. While the two head nouns are closely connected, in the real world the head noun that represents the place might be broader or at a larger distance in space.

Table 3

The quantiles of the Spearman rank correlation and the RSA metric computed considering the distance between objects in the image and the distance between their corresponding head nouns in the image caption, according to the scene tree.

Spearman Correlation			RSA		
Q1.	Q2	Q3	Q1.	Q2	Q3
0.59	0.9	0.99	0.0	0.69	0.89

Table 4

The quantiles of the Spearman rank correlation and the RSA metric computed considering the distance between objects in the image and the distance in the scene graph.

Spearman Correlation			RSA		
Q1.	Q2	Q3	Q1.	Q2	Q3
0.89	0.99	1.0	0.88	0.99	1.0

5.2.3. Correlations between the Visual Structure of the Image and the Scene Tree of the Caption

We create the scene tree by processing the parsed dependency tree (from the previous subsection) such that it only contains the head nouns matching with image objects. We compared the distances between the head nouns in the scene tree and between the objects in the image. The resulting distribution of correlations is shown in Table 3. Comparing with the results reported for the dependency and the constituency trees, we see that the correlation scores are higher for the case of scene trees. This might be due to the depth of the scene trees which is in general smaller than the depth of the dependency and the constituency trees. While for the RSA metric the first quantile is still zero, the median is 0.69 and the third quantile is almost 0.9. This indicates that half of the scene trees have a good correlation with distances in the image. For the Spearman rank correlations, the scores are even better, with a median of 0.9 and a third quantile of 0.99. This would mean that a quarter of the scene trees are perfectly correlated with the distances in the image.

5.2.4. Correlations between the Visual Structure of the Image and the Scene Graph that Describes the Image

Here we compare the distances between object nodes in the scene graph with distances in the corresponding image. Note that in the previous subsections, the head nouns in the caption decided on the object boxes in the image, while here scene graphs are not associated with captions but form their own language and alignment with object boxes in the image. We show the distribution of correlations of the computed distances in Table 4. We notice very high correlations with a median of 0.99 for both the RSA and the Spearman rank correlation. This is as expected since the relations of the visual genome are defined through the definition of smaller subregions indicating mostly just one relation (see Section 2.4). Therefore, objects far apart in the image, are probably not annotated together in a region.

6. Discussion and Open Questions

In our preliminary experiments, we looked at the correlations between object distances in different language structures and spatial distances in the image. The results for this are reported in Section 5. There is a clear difference in correlations between the different language structures used.

Both the well-established language grammar parses have very similar results and both show a positive correlation with distances in the image. This fact means that increasing the distance between two objects in the physical world is represented by a higher distance between the corresponding head nouns in the dependency and constituency trees. However, the correlations are moderate given that we can connect through language not only very short-distance objects but also object with a large spatial distance (i.e., objects that shows location). The language grammars are designed to bring structure to language such that common patterns are easy to understand when communicating. These structures have limitations in describing the world.

It is surprising that there is no difference in correlations between the dependency tree paradigm and the constituency tree paradigm. In Section 3 we discussed scene grammars and the possible structure humans impose on the world. These grammars seem to be hierarchical in nature and can be seen as constituency trees. However, this can be explained that the language constituency tree creates phrases for groups of words in the sentence close to each other in position, while placing nouns in different noun phrases. However, the scene grammar would group the objects (described by nouns) close to each other around an anchor object.

For the scene tree, we notice much higher correlations, even though it is a reduced version of the dependency tree to only the head nouns. However, note that in either case, we are only computing distances between the head nouns. In the dependency tree, the path between these nodes has to go through the entire sentence, possibly drastically changing distances, while in the scene trees, these paths are much more direct. We also notice a large difference in the distributions for RSA scores and Spearman rank correlation scores. 50% of the pairs have a Spearman rank correlation higher than 0.9. It is possible that the scene tree has become more flat compared to the dependency tree, making it more trivial. In this case, the RSA is probably more informative and a better measure of the correlations.

For the scene graphs, we notice that the distances between head nouns almost perfectly align with the distances in the image. While the scene graph is a language structure, it is constructed directly from observations from the image already creating an immediate correlation. As already mentioned, the annotations are created based on small regions, causing close objects to be directly mentioned in the regions, while further objects are only connected through multiple steps in the graph. Despite these results, the scene graph is not a perfect structure. On average, each scene graph has 35 objects, 26 attributes, and 21 pairwise relationships between objects per image. This is very dense and is caused by the many region descriptions. When looking at the example in Figure 3, we notice that many of the objects are already mentioned in the descriptions and directly connecting them, including many very specific details, i.e. all the details about the girls' feet such as the toes, the bottom of them, that they are bare, and that the toes are bare. These mentions can easily be derived through common sense reasoning and it could help to exclude them to have a more understandable graph. The number of descriptions and details results in many very short paths, which could partly explain the higher correlation scores.

Information in the scene graphs is often repeated. For example, in Figure 3, the laptops are separately described for each person, but also several times collectively with different sentence structures and details. Furthermore, because of the free-form text of the descriptions, the labels for all annotations are extremely sparse (despite mapping everything to their WordNet synsets [34]), and there is no rule on how the relations should be ordered. Take for example the relations between the shirt and the man in Figure 3, it both says that the man is wearing the shirt and that the shirt is on the man. This makes it more difficult to study common graph patterns and structures and contradicts the hierarchical nature of objects in scenes, hypothesizes for the scene grammar in Section 3.

Open questions For language grammar, extensive research on large corpora has properly established the different types of grammar and its rules. For other modalities, studies regarding grammar are only scratching the surface. Võ et al. [6] have found evidence that there are some structures in visual data that appear to be like a grammar. However, the studies are mainly coming from a psychological standpoint. This raises the question of whether similar clues can be found from a more data sci-

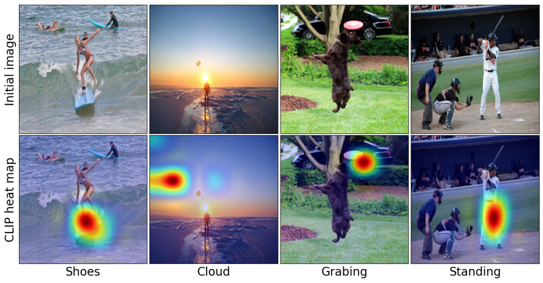

Figure 6: Example of image-language relations extracted with CLIP [10]. The heat map visualizations are computed by applying GRAD-CAM [57] over CLIP. Source images: COCO 2017 validation dataset.

ence standpoint. By extending such research to different modalities and designing new datasets and annotations, we can improve our understanding of how humans process scenes and the visual world, and how this processing correlates to our language.

In the current study, we measured correlations of object distances between language structures and visual structures. From these, it is still difficult to infer to what extent language structures have influenced the way humans see or organize the physical world, or vice versa. More research is here needed. Moreover, more fine-grained syntactical or semantic structures could be studied when inferring this influence.

How can such studies help to create better content representations used in artificial intelligence? The current visio-linguistic foundation models have already made substantial progress towards capturing even the most subtle relations between language and the physical world that go beyond only recognizing objects in the images. For example, the CLIP model [10] can identify the regions of the missing objects by relying on semantic concepts that connect the missing objects with the already available objects in the image, following the human way of thinking [58]. Considering Figure 6, CLIP identifies the region of shoes for a barefoot person and the region of clouds in the cloudless sky. The visio-linguistic foundation models, exemplified by CLIP in Figure 6, can also detect semantic but non-visual concepts like actions. When asking CLIP to detect the region of the verb “grabbing”, the model knows that the dog is grabbing a frisbee and the selected region is represented by the dog’s mouth and frisbee. Similarly, the region of the verb ”standing” is represented by the man’s legs. While these capabilities are impressive, they raise the question of what structures are learned. Has it learned a semantic and syntactic memory for positions and interactions of objects in scenes, similar to those in humans [49]? It could help in the discovery of a scene grammar by further investigating the capabilities of foundation models and how well their learned representation correlates to the human structuring of scenes.

7. Conclusion

In this work, we explored the commonalities between structures in language and in visual scenes. While finding some common structures and patterns in how humans process visual scenes and language, it is not evident to correlate language structure with image structures. With some preliminary experiments, we found some good correlations between some of the existing language structures and object locations in images.

With these experiments and the created overview of existing structures, we were able to state some of the open questions. We hope that this overview can lead to better collaborations between the different fields of natural language processing, computer vision, and psychology aiming to better understand how the world is structured and the influence of language and of the physical world on that structure. Such studies might pave the way to improved representations of language and visual content that take into account structural knowledge.

Acknowledgments

Victor Milewski, Maria Trusca, and Marie-Francine Moens were funded by the European Research Council (ERC) Advanced Grant CALCULUS (grant agreement No. 788506).

References

[1] G. Ganis, M. Kutas, An electrophysiological study of scene effects on object identification, Cognitive Brain Research 16 (2003) 123–144.

[2] L. Mudrik, D. Lamy, L. Y. Deouell, Erp evidence for context congruity effects during simultaneous object–scene processing, Neuropsychologia 48 (2010) 507–517.

[3] M. L.-H. Võ, J. M. Wolfe, Differential electrophysiological signatures of semantic and syntactic scene processing, Psychological Science 24 (2013) 1816–1823. URL: https://doi.org/10.1177/0956797613476955. doi:10.1177/0956797613476955, pMID: 23842954.

[4] L. Mudrik, S. Shalgi, D. Lamy, L. Y. Deouell, Synchronous contextual irregularities affect early scene processing: Replication and extension, Neuropsychologia 56 (2014) 447–458.

[5] N. Cohn, R. Jackendoff, P. J. Holcomb, G. R. Kuperberg, The grammar of visual nar-

rative: Neural evidence for constituent structure in sequential image comprehension, Neuropsychologia 64 (2014) 63–70. URL: https://www.sciencedirect.com/science/article/pii/S0028393214003236. doi:https://doi.org/10.1016/j.neuropsychologia.2014.09.018.

[6] M. L.-H. Võ, S. E. Boettcher, D. Draschkow, Reading scenes: how scene grammar guides attention and aids perception in real-world environments, Current Opinion in Psychology 29 (2019) 205–210. URL: https://www.sciencedirect.com/science/article/pii/S2352250X18302574. doi:https://doi.org/10.1016/j.copsyc.2019.03.009, attention & Perception.

[7] V. Cook, M. Newson, Chomsky's universal grammar: An introduction, John Wiley & Sons, 2014.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer, 2020, pp. 104–120.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] D. Jurafsky, J. H. Martin, Speech and language processing (draft), 3rd, 2019.

[13] N. Chomsky, Three models for the description of language, IRE Transactions on information theory 2 (1956) 113–124.

[14] W. M. Wundt, Völkerpsychologie; eine untersuchung der entwicklungsgesetze von sprache, mythus und sitte, volume Band II: Die Sprache, Zweiter Teil, W. Engelmann, 1900.

[15] D. McClosky, E. Charniak, M. Johnson, Effective self-training for parsing, in: North American Chapter of the Association for Computational Linguistics, 2006, pp. 152–159.

[16] C. Dyer, A. Kuncoro, M. Ballesteros, N. A. Smith, Recurrent neural network grammars, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 199–209. URL: https://aclanthology.org/N16-1024. doi:10.18653/v1/N16-1024.

[17] Y. Zhang, H. Zhou, Z. Li, Fast and accurate neural crf constituency parsing, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4046–4053. URL: https://doi.org/10.24963/ijcai.2020/560. doi:10.24963/ijcai.2020/560, main track.

[18] Y. Tian, Y. Song, F. Xia, T. Zhang, Improving constituency parsing with span attention, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1691–1703. URL: https://aclanthology.org/2020.findings-emnlp.153. doi:10.18653/v1/2020.findings-emnlp.153.

[19] L. Tesnière, Eléments de Syntaxe Structurale, Klincksieck, Paris, 1959.

[20] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al., Universal dependencies v1: A multilingual treebank collection, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 1659–1666.

[21] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://aclanthology.org/2021.cl-2.11. doi:10.1162/coli_a_00402.

[22] J. Nivre, An efficient algorithm for projective dependency parsing, in: Proceedings of the eighth international conference on parsing technologies, 2003, pp. 149–160.

[23] E. Kiperwasser, Y. Goldberg, Simple and accurate dependency parsing using bidirectional LSTM feature representations, Transactions of the Association for Computational Linguistics 4 (2016) 313–327. URL: https://aclanthology.org/Q16-1023. doi:10.1162/tacl_a_00101.

[24] T. Ji, Y. Wu, M. Lan, Graph-based dependency parsing with graph neural networks, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2475–2485. URL: https://aclanthology.org/P19-1237. doi:10.18653/v1/P19-1237.

[25] K. Mrini, F. Dernoncourt, Q. H. Tran, T. Bui, W. Chang, N. Nakashole, Rethinking self-attention: Towards interpretability in neural parsing, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computa-

tional Linguistics, Online, 2020, pp. 731–742. URL: https://aclanthology.org/2020.findings-emnlp.65. doi:10.18653/v1/2020.findings-emnlp.65.

[26] J. Hewitt, C. D. Manning, A structural probe for finding syntax in word representations, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4129–4138. URL: https://aclanthology.org/N19-1419. doi:10.18653/v1/N19-1419.

[27] V. Milewski, M. de Lhoneux, M. Moens, Finding structural knowledge in multimodal-bert, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2022, pp. 5658–5671.

[28] L. A. Hendricks, A. Nematzadeh, Probing image-language transformers for verb understanding, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 3635–3644.

[29] E. Bugliarello, A. Nematzadeh, L. A. Hendricks, Weakly-supervised learning of visual relations in multimodal pretraining, arXiv preprint arXiv:2305.14281 (2023).

[30] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5238–5248.

[31] F. Liu, G. Emerson, N. Collier, Visual spatial reasoning, Transactions of the Association for Computational Linguistics 11 (2023) 635–651.

[32] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, A. Gatt, VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8253–8280. URL: https://aclanthology.org/2022.acl-long.567. doi:10.18653/v1/2022.acl-long.567.

[33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (2016) 32–73.

[34] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[35] D. Xu, Y. Zhu, C. B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5410–5419.

[36] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 5831–5840.

[37] K. Tang, H. Zhang, B. Wu, W. Luo, W. Liu, Learning to compose dynamic tree structures for visual contexts, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6619–6628.

[38] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3716–3725.

[39] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, A. G. Hauptmann, A comprehensive survey of scene graphs: Generation and application, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1. doi:10.1109/TPAMI.2021.3137605.

[40] Y. Zhong, J. Shi, J. Yang, C. Xu, Y. Li, Learning to generate scene graph from natural language supervision, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 1803–1814.

[41] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, W.-Y. Ma, Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6609–6618.

[42] S. Schuster, R. Krishna, A. X. Chang, L. Fei-Fei, C. D. Manning, Generating semantically precise scene graphs from textual descriptions for improved image retrieval, in: Proceedings of the Fourth Workshop on Vision and Language, Association for Computational Linguistics, 2015, pp. 70–80.

[43] J. M. Wolfe, M. L.-H. Võ, K. K. Evans, M. R. Greene, Visual search in scenes involves selective and nonselective pathways, Trends in cognitive sciences 15 (2011) 77–84.

[44] D. Draschkow, J. M. Wolfe, M. L.-H. Vo, Seek and you shall remember: Scene semantics interact with visual search to build better memories, Journal of Vision 14 (2014) 10–10.

[45] M. Le-Hoa Võ, J. M. Wolfe, The role of memory for visual search in scenes, Annals of the New York Academy of Sciences 1339 (2015) 72–81.

[46] Y. Hong, Q. Li, S.-C. Zhu, S. Huang, Vlgrammar: Grounded grammar induction of vision and language, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1665–1674.

[47] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, S. Fidler, Meta-sim: Learning to generate synthetic datasets, in:

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4551–4560.

[48] J. Devaranjan, A. Kar, S. Fidler, Meta-sim2: Unsupervised learning of scene structure for synthetic data generation, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 715–733.

[49] D. Draschkow, M. L.-H. Võ, Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search, Scientific reports 7 (2017) 16471.

[50] N. Cohn, Explaining 'i can't draw': Parallels between the structure and development of language and drawing, Human Development 55 (2012) 167–192.

[51] R. Jackendoff, What is the human language faculty? two views, Language (2011) 586–624.

[52] R. S. Jackendoff, Language, consciousness, culture: Essays on mental structure, MIT Press, 2009.

[53] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.

[54] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, International Journal of Computer Vision 123 (2015) 74–93.

[55] N. Kitaev, D. Klein, Constituency parsing with a self-attentive encoder, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2676–2686.

[56] B. P. Kriegeskorte Nikolaus, Mur Marieke, Representational similarity analysis - connecting the branches of systems neuroscience, Frontiers in Systems Neuroscience (2008).

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, IEEE Computer Society, 2017, pp. 618–626.

[58] M. L.-H. Vo, The meaning and structure of scenes, Vision Research 181 (2021) 10–20.