Measuring Perceptual and Linguistic Complexity in Multilingual Grounded Language Data

Nisha Pillai
Univ. of Maryland, Baltimore County
npillai1@umbc.edu

Cynthia Matuszek
Univ. of Maryland, Baltimore County
cmat@umbc.edu

Francis Ferraro
Univ. of Maryland, Baltimore County
ferraro@umbc.edu

Abstract

The success of grounded language acquisition using perceptual data (*e.g.*, in robotics) is affected by the complexity of both the perceptual concepts being learned, and the language describing those concepts. We present methods for analyzing this complexity, using both visual features and entropy-based evaluation of sentences. Our work illuminates core, quantifiable statistical differences in how language is used to describe different traits of objects, and the visual representation of those objects. The methods we use provide an additional analytical tool for research in perceptual language learning.

Introduction

In *grounded language acquisition,* examples from a physical or simulated context are used to drive language learning. While there has been significant recent effort on grounded language learning (Huang et al. 2019; Thomason et al. 2019b; Wu et al. 2019; Zhou, Arnold, and Yu 2019), handling grounded language remains a challenging problem, in part because groundings are learned from two noisy, ambiguous, and complex channels. We provide a mechanism for analyzing the visual and linguistic complexity of the data used, so as to better understand the trait modifiers, such as color and shape, of the inputs involved.

Prior studies in multimodal grounding (Kery et al. 2019; Pillai et al. 2020) show that the amount of data required to learn about different traits of an item—such as its color, shape, or overall object type—varies significantly, with speculation that the "complexity," broadly defined, of the trait being linguistically described (or visually represented) is a key correlate to this varied performance. As in Fig. 1, this is fairly intuitive, though the lack of a clear quantitative measure limits the conclusions that can be drawn.

This work is a focused effort to demystify and quantify this "complexity." We propose straightforward, approachable measures for computing both visual and linguistic complexity. We use two existing datasets (Lai 2014; Pillai and Matuszek 2018) that contain RGB+depth images of a variety of objects labeled in multiple languages (Kery et al. 2019), and we consider language describing three different conceptual types: "color," "shape," and "object" (Fig. 1). We compute statistical descriptors with these calculated complexities, and analyze how both the complexity scores and statistical descriptors align with broad human intuition.

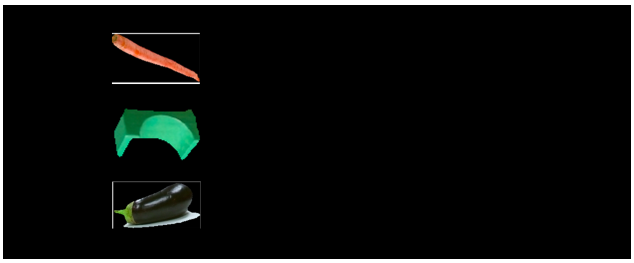

Figure 1: RGB-D sensor data and descriptions. Colors are described simply and fairly consistently. Shapes are a more difficult problem visually, and tend towards noisy descriptions. Object types are linguistically more consistent, but are the most difficult perceptual problem, in part due to the specificity of labels.

Our primary contributions are (1) the introduction of "trait-based" complexity to the AI and grounded language learning communities, and (2) the identification of appropriate metrics and statistical tools to measure the complexity of perceptual data and linguistic variety to predict efficiency in grounded language learning. Language is measured using a sentence-based entropy analysis, and visual complexity by examining visual featurizations. We argue that this straightforwardness of our approach is a benefit, since it is simple to compute yet effective at discerning key differences in grounded language. We further argue that the complexity measures provide grounded language learning researchers with an additional tool for analyzing and understanding their data and underlying learning problem.

Related Work on Grounding Language and Quantifying "Complexity"

Grounded language acquisition in perceptual data is related to a wide range of other problems, and our complexity evaluation has the potential to be relevant to many of them. We foresee applications to areas including visual navigation (Nguyen et al. 2019; Jain et al. 2019; Hu et al. 2019;

Nguyen and Daumé 2019); scene generation (Cheng et al. 2019); understanding videos (He et al. 2019; Antol et al. 2015); image captioning (Song et al. 2019; Jiang et al. 2019); visual story telling (Huang et al. 2016); and object detection (Huang, Chang, and Hauptmann 2019). Grounded object description, where a single object is described, is most closely related to this effort (Richards and Matuszek 2019).[1]

We expect our approach to generalize to many of the language grounding problems that are currently of significant interest to the field (Thomason et al. 2019a; Patki et al. 2019; Goyal, Niekum, and Mooney 2019; Chai et al. 2018; Liang et al. 2018, i.a.). Despite our focus on RGB-D data, variations of our measures should apply to most data with a visual component; the language analyses are directly applicable. There is a growing interest in multilingual grounded language. These efforts encompass both image captioning (Elliott et al. 2016; Hewitt et al. 2018, i.a.), learning spatial relations (Belz et al. 2018; Elliott and Keller 2013), and multilingual grounded object description (Kery et al. 2019).

Linguistic complexity has been investigated via numerous psycholinguistic approaches, including concreteness and imageability (Shi et al. 2019; Hessel, Mimno, and Lee 2018, i.a.), cost of learning (Becerra-Bonache, Christiansen, and Jiménez-López 2018), and length of words (Lewis and Frank 2016). While Ferraro et al. (2015) presented multiple syntactic-, concreteness-, and language modeling-based approaches for quantifying the complexity of vision-and-language-based *datasets*, we are interested in examining the complexity of semantic *traits* (categories of concepts) encompassed by those datasets. Many of these are less relevant to the problem we study, in that we intend to quantify the differences in complexity, rather than discover the cognitive sources of those differences. Further, as our data is drawn from robotics, almost all of the concepts being learned are similarly concrete (vs. abstract).

Computationally measuring visual complexity in accordance with human perception is challenging. Human reactions can be influenced by familiarity, style, and other perceived factors, which is challenging to evaluate (Machado et al. 2015). However, here we intend to find an automated measure for the *concept complexity* of an image (Miniukovich and De Angeli 2016). As shape complexity varies widely (Attneave 1954), for this category we use compression techniques (Forsythe, Mulhern, and Sawey 2008; Donderi 2006).

Approach: Measuring Complexity

Although perceptual and linguistic complexity are intuitive concepts to people, they are difficult to verbalize or define. In general, humans are poor at providing numerical priors or rankings of subtle concepts, particularly over a very large dataset (Aroyo and Welty 2015). We accordingly introduce automated metrics here. This paper is an attempt to clarify the concept of visual and linguistic complexity individually. Approximating combined visual-linguistic complexity would be an exciting topic for future study.

In identifying these metrics, we do not claim that they are the only possible metrics. Indeed, we argue that the complexity measures provide grounded language learning researchers with an additional tool for analyzing and understanding their data and underlying learning problem. We hope that our work will encourage the community to begin to examine these notions of complexity in their other efforts and for other grounded language tasks.

Datasets

In order to limit any confounding task-oriented aspects, we analyze two datasets that are used for concept grounding. We chose these datasets as they have an appropriate mixture of color, shape, and object descriptions.

The first is the well-known UW RGB-D+ object set (Lai et al. 2011), which contains images for 300 objects across 51 categories, annotated with 1,500 human-provided English descriptions (Richards and Matuszek 2019). Second, we use the UMBC RGB-D object/language dataset (Pillai and Matuszek 2018), which has 72 objects divided into 18 classes. Each object has multiple descriptions, for a total of 6,000 English, 5,100 Spanish and 5,700 Hindi descriptions (Kery et al. 2019). Items include food objects such as 'tomato' and 'corn' and blocks in shapes such as 'cube' and 'triangle.' These datasets contain an average of 4.9 (UW RGB-D+) and 4.5 (UMBC) images per object. While the datasets we use are not very large, there are comparatively few available datasets that combine *non-simulated, robotic-inspired* data (in contrast to, e.g., Scalise et al. (2019)) collected using modern RGB+depth robotic sensors (as opposed to, e.g., ImageNet) and natural language descriptions.

Previous work has successfully used a "concepts-as-classifiers" method for grounded language learning (Schlangen, Zarrieß, and Kennington 2016; Pillai et al. 2020): visual classifiers are learned that directly associate visual features with a lexically-oriented concept word. These concepts are extracted from descriptions via tokenization and stopword removal. For example, an image of a tomato described as "This looks like a red tomato without any leaves on the top," yields the *concepts* RED, TOMATO, and LEAVES.

Linguistic Complexity

We calculate the linguistic complexity by computing lexical entropy, extracted for each concept from the descriptions. For every object instance i, we combine all the descriptions into a pseudo-document d_i. We calculate the frequency for every descriptive concept v in d_i as $p_{i,v} \propto \text{count}(v \in d_i)$. We then compute the entropy h_i of the instance i: $h_i = -\sum_u p_{i,u} \log p_{i,u}$. Descriptions and entropies can be separated at the pseudo-document level depending on whether a characteristic word, e.g., "red," was used or not. As the entropy reflects the diversity of language used to describe an instance, examining its variability helps explain the linguistic complexity of a trait. This approach, though straightforward, is nicely consistent with the traditional concepts-as-classifier grounding approach used in

[1] In contrast to most image captioning tasks, grounded object description allows people to describe how, or why, an item can be used in particular ways. It focuses on the *object* rather than the *image* containing the object.

Data-set	Language	*D*, Color Concepts	*D*, Shape Concepts	*D*, Object Concepts
UMBC	English	0.41 (1.63E-7)	0.28 (1.163E-3)	0.36 (9.0E-6)
	Spanish	0.58 (1.3E-14)	0.23 (1.3E-2)	0.39 (9.9E-7)
	Hindi	0.48 (5.4E-10)	0.41 (2.0E-7)	0.58 (2.8E-14)
UW RGB-D+	English	0.20 (2.2E-16)	0.56 (2.2E-16)	0.56 (2.2E-16)

Table 1: Kolmogorov-Smirnov test result of each dataset and language on trait *vs.* not-trait. *D* represents the max distance between the two samples' empirical CDF, i.e., the trait and non-trait cumulative distributions. All results are significant to at least p=0.013, with p-values provided in parentheses. This table shows that the UMBC dataset has fairly consistent color descriptions (larger K-S distances), but the UW-RGBD dataset—which contains more complex, multi-colored objects—is less consistent (smaller distances). K-S distances for shape and object traits are smaller, indicating complex, varied descriptions.

previous work (Schlangen, Zarrieß, and Kennington 2016; Thomason et al. 2016; Abend et al. 2017).

We then calculate the density estimates of entropy for the distribution of language describing a trait such as color vs. the distribution of language not describing that trait. Descriptions of one instance may include concepts of all three traits (e.g., "a round purple eggplant"). We then calculate the linguistic complexity of a trait by combining the entropies of descriptions that reference a concept associated with that trait. We compare them with the cumulative entropy calculated from all other concepts that are not related to that trait. For example, we combine all the descriptions of eggplant instances and calculate the entropy using every concept's count. For color, we select all the concepts associated with color (e.g., "purple"), and add the entropies of all the instances described by that concept ("purple"). To calculate the distribution of non-color traits, we add all the entropies which are not color. We categorized the concepts corresponding to each trait with the help of Google translate (Wu et al. 2016).[2]

We perform a Kolmogorov-Smirnov (K-S) test to quantify these distributions. K-S tests are an efficiently compare two distributions or samples against the null hypothesis that they do not differ. The K-S test returns the maximum distance *D* between two curves, with *D* bounded by 0 (for identical distributions) and 1. Results are shown in Tab. 1. This quantifies the "difference" between two distributions. A K-S test also provides an efficient way to reject a null hypothesis (the two distributions do not differ).

Visual Complexity

To estimate the complexity and variability of visual traits, We use edge density features and compression errors, as proposed and validated by Machado et al. (2015). We consider the different categories in concept-specific ways. For color, the approach is simple: we use the empirically validated approach of computing the standard deviation of RGB values.

[2]Available with code at https://github.com/iral-lab/MultiModalComplexityEval.

To measure shape complexity, we compute the compression loss of detected edges. We extract HSV values from an RGB image, compute edge densities over these using standard edge detection algorithms (Canny; Kanopoulos, Vasanthavada, and Baker (1986; 1988)), and estimate the compression errors using JPEG compression (see Fig. 3). Machado et al. (2015) presented user studies validating this approach; other compression techniques would need to be validated in a similar fashion—an effort deserving of a dedicated study.[3]

In order to meaningfully analyze object type as distinct from color and shape, we would need to consider a different featurization that captured more of the semantics of "object-ness" (Pillai and Matuszek 2018; Kery et al. 2019). This is a topic for future work. Nevertheless, we expect our approach to generalize beyond the language grounding problems that are currently of significant interest to the field (Goyal, Niekum, and Mooney 2019; Anderson et al. 2018). We focus on RGB-D data, but variations of our measures apply to most data with a visual component, while our language analyses are directly applicable.

Analysis of Linguistic Complexity

Fig. 2 shows the density computed from UMBC and UW RGB-D+ datasets' entropies. The variability of the traits can be seen from the entropy results in the figure.

English: We can see that color entropies are concentrated towards zero compared to non-color entropies, indicating the concise, less diverse vocabulary used to describe colors. For example, the BLUE concept is described using exactly the term "blue" in 95% of the descriptions. Non-color entropies are more diverse, indicating the variance in the descriptions; "color" is linguistically simpler in these datasets than other traits. "Shape" is the most varied trait, with high variance in the annotations, both according to our metrics and in practice. "Object" annotations are more consistent than "shape" as users were mostly consistent in describing vegetables ("corn," "cabbage") but less consistent in annotating toys ("arch," "cube").

Hindi: We see that color complexity (that is, diversity of language describing color) is much smaller than that of shape and object. From the annotations, we find the different forms of the same words are used to describe the object: The "color" concepts are semantically similar but exhibit noun inflection based on gender. Such discrepancies affect language acquisition performance. Diverse words are used for shapes, particularly to describe cylindrical objects, making the downstream language learning problem more complex. High entropy implies weak agreement between the annotators. The patterns of complexity among traits in Hindi are nonetheless approximately analogous to English.

Spanish: The diversity of terms used in Spanish across the three traits is similar to that of English and Hindi: language

[3]Canny edge detection coupled with JPEG compression provides one of the highest correlations between human and computational estimates of visual complexity. This implies that edge density and compression error are reliable predictors of people's perception of visual complexity.

Dataset	Lang.	Color	Shape	Object Type
UMBC	**English**	Concise, less varied concept vocabulary	Contrasting descriptions; 83% of instance descriptions included shape	Varied and diverse descriptions; all instances are described with object names
	Spanish	Concise, less varied concept vocabulary	Many synonyms; all instances are described with shape concepts at least once	Varied and diverse descriptions; all instances are described with object types at least once
	Hindi	Semantically similar, but gender-based inflectional differences present	Highly varied and diverse concepts; all instances are described with shape concepts	Varied and diverse descriptions; all instances are described with object concepts at least once
UW RGB-D+	**English**	Multicolor objects, medium consistency in descriptions; not all descriptions include color	Synonyms present; only 9.5% of all instances have shape in descriptions	Varied and diverse descriptions; 98.9% instances have object concepts in description

Table 2: A qualitative summary of the typical complexity of linguistic descriptions by dataset and language.

Data-set	Language	Color		Shape		Object	
		Yes	No	Yes	No	Yes	No
UMBC	**English**	0.71	1.45	1.20	1.18	1.67	0.95
	Spanish	1.17	2.23	2.07	1.78	2.38	1.63
	Hindi	1.09	1.61	0.95	1.68	2.27	1.03
UW RGB-D+	**English**	0.39	0.58	0.01	0.71	1.03	0.22

Table 3: The average of linguistic complexity comparisons between trait vs. non-trait for each dataset and language. Higher differences between average values indicates the conciseness in the description: color descriptions are concise compared to shape and object descriptions.

Dataset	Color	Shape
UMBC	0.120	0.910
UW RGB-D+	0.171	0.942

Table 4: The average value of visual complexity measures of color and shape distributions for every dataset. The smaller mean for our color complexity metric indicates a lack of variety in color features, while larger values for shape complexity are a result of the complicated edges and shapes in the feature set.

about colors is consistent and straightforward, but becomes more complex for shape and object. "Color" shows the least variation of the three traits, although there is more variance in color descriptions for concepts with similar meanings, such as the very similar terms *morado*, *púrpura*, and *violeta* for purple. The vocabulary used for shape features is varied and inconsistent. All of *rectangulo*, *poliedro*, and *paralelepípedo* appear when describing rectangular solids. Similarly, object terms vary widely, possibly due to a difference in what objects are routinely found and discussed in day-to-day life. For example, a cucumber was described as a *pepino* (cucumber) and *pepinillo* (pickle), but also several times as "looking like a small *sandía* (watermelon)," as well as by the category hypernyms *vegetal* and *fruta* (vegetable and fruit).

Overall, the relative linguistic complexity of traits is comparable to that of English and Hindi. All three languages have a consistent and straightforward vocabulary for the "color" concept, but varied and complex vocabulary for "shape" and "object" concepts.

There are differences between the datasets. In the UW RGB-D+ dataset, not every instance is described with a color, which is reflected in the lower K-S distances. "Object" descriptions in the UW dataset are also more diverse compared to the UMBC dataset. The atypical "shape" behavior indicates the lack of "shape" words: only 9.5% of instance descriptions have shape descriptors, likely due to the occurrence of more complex objects with less of simple geometric shape. In summary, while language-specific differences do emerge, we see very similar overall patterns of complexity across languages.

Analysis of Visual Complexity

In modeling visual complexity, we consider shape and color differences between the two datasets, omitting object type for the reasons described above. Results are shown in Fig. 3.

In the smaller UMBC dataset, the standard deviations of RGB values are a good indication of greater visual consistency, while lower compression rates are a good indication of reduced complexity. From these results, we can conclude that the overall color deviation is small, which is accurate for the dataset being measured. The compression rates of shape concepts are more varied, indicative of greater visual variety.

Results are similar in the UW RGB-D+ dataset. While there are subtle differences, the overall complexity profile between the two datasets is similar. There is more diversity in the UW dataset color standard deviation, presumably due to the less monochromatic objects in this dataset.

Previous work reported large performance drops in classification surrounding "color"-concepts vs. "shape"-based concepts (Pillai and Matuszek 2018); while "color" yielded an accuracy of 0.81, "shape" was much lower at 0.62. This roughly tracks with our complexity measures: both linguistic and visual complexity measures for the "color" trait are lower (indicating lower complexity, and more successful classification) while the complexity measures for the "shape" are higher (indicating higher complexity, more complicated descriptions/visuals, and harder classification). Additionally, we find that in the context of dealing with concrete objects, the level of ambiguity in learning varies with multi-sense concepts. For instance, "orange" is both a color and an object. Learning the meaning of "orange" as a color is simpler than "orange" as an object, and our complexity measures reflect that.

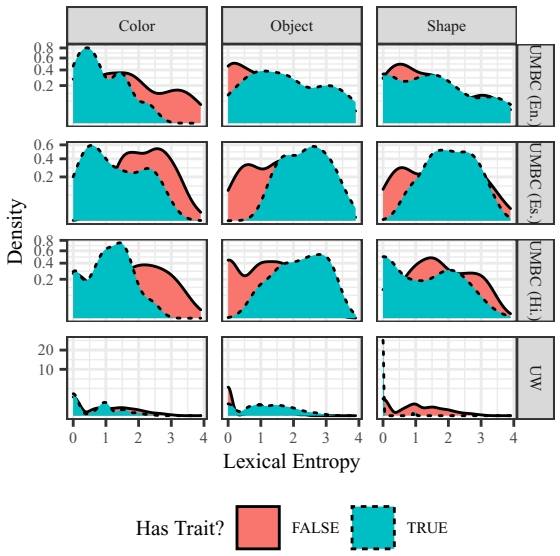

Figure 2: Comparison of traits "Color," "Shape," and "Object" via lexical entropy for the UMBC and UW RGB-D+ datasets. The K-S statistics quantify the amount of divergence within each facet (subplot). Notice the entropy for color concepts is lower than non-color, indicating the concise, less varied vocabulary of colors. Object trait entropy is higher, indicating linguistic variability. Only 9.5% of the UW dataset instances have shape concepts in the description at least once. Spanish descriptions contain varied but semantically similar shape/object tokens in vocabulary.

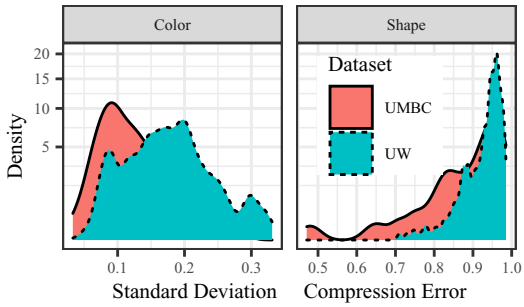

Figure 3: Visual complexity of "color" & "shape" for both datasets. Lower standard deviations are a good indication of greater visual color consistency. The left-skew of the compression errors illustrates the high variations of "shape."

Conclusion

In this work, we analyzed multilingual grounded language data and presented models that allow automated analysis of the complexity of descriptions and visual inputs. We verify that there is a consistent, statistically verifiable pattern of complexity across the traits we consider, making it possible to consider differentiated learning approaches in the cross-modal grounding tasks. We anticipate this will help grounded language learning researchers better understand the data they are working with, and yield and aid improved design decisions, such as appropriate feature selection and selection of classification model.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant Nos. 1657469, 1940931, and 2024878.

References

Abend, O.; Kwiatkowski, T.; Smith, N.; Goldwater, S.; and Steedman, M. 2017. Bootstrapping language acquisition. *Cognition*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and van den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: visual question answering. In *ICCV*.

Aroyo, L., and Welty, C. 2015. Truth is a lie: crowd truth and the seven myths of human annotation. *AI Magazine*.

Attneave, F. 1954. Some informational aspects of visual perception. *Psychological review*.

Becerra-Bonache, L.; Christiansen, H.; and Jiménez-López, M. D. 2018. A gold standard to measure relative linguistic complexity with a grounded language learning model. In *COLING Workshop on LC&NLP*.

Belz, A.; Muscat, A.; Anguill, P.; Sow, M.; Vincent, G.; and Zinessabah, Y. 2018. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *INLG*.

Canny, J. 1986. A computational approach to edge detection. *TPAMI*.

Chai, J. Y.; Gao, Q.; She, L.; Yang, S.; Saba-Sadiya, S.; and Xu, G. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*.

Cheng, Y.; Shi, Y.; Sun, Z.; Feng, D.; and Dong, L. 2019. An interactive scene generation using natural language. In *ICRA*.

Donderi, D. 2006. Visual complexity: A review. *Psychological Bulletin*.

Elliott, D., and Keller, F. 2013. Image description using visual dependency representations. In *EMNLP*.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30k: Multilingual english-german image descriptions. *ArXiv*.

Ferraro, F.; Mostafazadeh, N.; Huang, T.-H. K.; Vanderwende, L.; Devlin, J.; Galley, M.; and Mitchell, M. 2015. A survey of current datasets for vision and language research. In *EMNLP*.

Forsythe, A.; Mulhern, G.; and Sawey, M. 2008. Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*.

Goyal, P.; Niekum, S.; and Mooney, R. J. 2019. Using nat-

ural language for reward shaping in reinforcement learning. In *IJCAI*.

He, D.; Zhao, X.; Huang, J.; Li, F.; Liu, X.; and Wen, S. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*.

Hessel, J.; Mimno, D.; and Lee, L. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*.

Hewitt, J.; Ippolito, D.; Callahan, B.; Kriz, R.; Wijaya, D. T.; and Callison-Burch, C. 2018. Learning translations via images with a massively multilingual image dataset. In *ACL*.

Hu, R.; Fried, D.; Rohrbach, A.; Klein, D.; Darrell, T.; and Saenko, K. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *ACL*.

Huang, T.-H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R. B.; He, X.; Kohli, P.; Batra, D.; Zitnick, C. L.; Parikh, D.; Vanderwende, L.; Galley, M.; and Mitchell, M. 2016. Visual storytelling. In *HLT-NAACL*.

Huang, B.; Bayazit, D.; Ullman, D.; Gopalan, N.; and Tellex, S. 2019. Flight, camera, action! using natural language and mixed reality to control a drone. *ICRA*.

Huang, P.-Y.; Chang, X.; and Hauptmann, A. G. 2019. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *EMNLP*.

Jain, V.; Magalhães, G.; Ku, A.; Vaswani, A.; Ie, E.; and Baldridge, J. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*.

Jiang, M.; Huang, Q.; Zhang, L.; Wang, X.; Zhang, P.; Gan, Z.; Diesner, J.; and Gao, J. 2019. Tiger: Text-to-image grounding for image caption evaluation. *ArXiv*.

Kanopoulos, N.; Vasanthavada, N.; and Baker, R. L. 1988. Design of an image edge detection filter using the sobel operator. *JSSC*.

Kery, C.; Pillai, N.; Matuszek, C.; and Ferraro, F. 2019. Building language-agnostic grounded language learning systems. In *Ro-Man*.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2011. A large-scale hierarchical multi-view RGB-D object dataset. In *ICRA*.

Lai, K. K. W. 2014. *Object Recognition and Semantic Scene Labeling for RGB-D Data*. Ph.D. Dissertation, University of Washington.

Lewis, M. L., and Frank, M. C. 2016. The length of words reflects their conceptual complexity. *Cognition*.

Liang, H.; Wang, H.; Wang, J.; You, S.; Sun, Z.; Wei, J.-M.; and Yang, Z. 2018. JTAV: Jointly learning social media content representation by fusing textual, acoustic, and visual features. In *COLING*.

Machado, P.; Romero, J.; Nadal, M.; Santos, A.; Correia, J.; and Carballal, A. 2015. Computerized measures of visual complexity. *Acta psychologica*.

Miniukovich, A., and De Angeli, A. 2016. Pick me!: Getting noticed on google play. In *CHI*. ACM.

Nguyen, K., and Daumé, H. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *ArXiv*.

Nguyen, K.; Dey, D.; Brockett, C.; and Dolan, B. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *CVPR*.

Patki, S.; Daniele, A. F.; Walter, M. R.; and Howard, T. M. 2019. Inferring compact representations for efficient natural language understanding of robot instructions. In *ICRA*.

Pillai, N., and Matuszek, C. 2018. Unsupervised selection of negative examples for grounded language learning. In *AAAI*.

Pillai, N.; Raff, E.; Ferraro, F.; and Matuszek, C. 2020. Sampling approach matters: Active learning for robotic language acquisition. *IEEE Big Data (special session on machine learning in Big Data)*.

Richards, L. E., and Matuszek, C. 2019. Learning to understand non-categorical physical language for human-robot interactions. In *R:SS workshop on AI+ACR*.

Scalise, R.; Thomason, J.; Bisk, Y.; and Srinivasa, S. 2019. Improving robot success detection using static object data. In *IROS*.

Schlangen, D.; Zarrieß, S.; and Kennington, C. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *ACL*.

Shi, H.; Mao, J.; Gimpel, K.; and Livescu, K. 2019. Visually grounded neural syntax acquisition. In *ACL*.

Song, L.; Liu, J.; Qian, B.; and Chen, Y. 2019. Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. In *AAAI*.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. J. 2016. Learning multi-modal grounded linguistic semantics by playing i spy. In *IJCAI*. AAAI Press.

Thomason, J.; Padmakumar, A.; Sinapov, J.; Walker, N.; Jiang, Y.; Yedidsion, H.; Hart, J.; Stone, P.; and Mooney, R. J. 2019a. Improving grounded natural language understanding through human-robot dialog. In *ICRA*.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2019b. Vision-and-dialog navigation. *CoRR*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*.

Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W.-Y. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*.

Zhou, M.; Arnold, J.; and Yu, Z. 2019. Building task-oriented visual dialog systems through alternative optimization between dialog policy and language generation. In *EMNLP*.