Capturing stage-level and individual-level information from photographs: Human-AI comparison

Yuri Sato (sato.yuri@ocha.ac.jp)
Department of Humanities Data Engineering & Department of Philosophy, Ochanomizu University, Japan

Ayaka Suzuki (suzuki.ayaka@chiba-u.jp)
Department of Japanese Language and Culture, Chiba University, Japan

Koji Mineshima (minesima@abelard.flet.keio.ac.jp)
Department of Philosophy, Keio University, Japan

Abstract

This study explores human capabilities in distinguishing and recognizing entities that change over time from those that do not. We specifically investigate the linguistic distinction between "individual-level predicates" (ILPs) and "stage-level predicates" (SLPs). Our empirical approach focuses on how humans visually distinguish these two types. We performed a corpus analysis, in which a limited set of image captions were randomly extracted and annotated by experts with either SLP or ILP labels. The findings indicated a predominance of SLPs over ILPs in the image captions, alongside identifying frequently occurring verbs associated with each type of predicate. Building on this manual annotation, we extended the process to automatic annotation on a broader dataset of image captions. This facilitated a machine-learning experiment for image classification based on ILPs and SLPs. Our results demonstrated that the classification of SLPs achieved a substantially high accuracy rate, though not as high as human accuracy, while the classification of ILPs had an accuracy rate of about chance level, substantially lower than human capabilities. Given the analyses, we discuss what features of the image contribute to distinguishing between ILPs and SLPs.

Keywords: image; image caption; individual-level; stage-level; machine learning; visual grounding

Introduction

It is a fundamental human ability to distinguish and recognize entities in the real world that change over time and those that do not. For instance, imagine you are in an unfamiliar town and you have memorized the route from the train station to a concert hall for your outward journey, intending to use the route for your return. Certain landmarks, like a magnificent church, several large ginkgo trees, or a distinctive graffiti on a wall, can serve as useful navigational clues. These features typically remain unchanged over short periods (such as the duration of a concert), making them reliable reference points since they do not move or disappear. In contrast, transient elements like a vintage red car parked on the street or a cat walking down the road are unlikely to be present in the same way on the return trip. Since such things are likely to change location over a short period of time, they are not normally used as landmarks in route memorization.

The distinction is not simply a matter of whether they are stationary or moving, but whether they have the *potential* to move. This kind of distinction has been made in philosophy, especially Ontology, since Aristotle. According to Zemach (1970) and Arp, Smith, and Spear (2015), anything that exists in the world is called "entities," and the first criterion for distinguishing them is whether or not they persist through time. If they do, they are assigned to the category of "continuants"; if not, they are assigned to the category of "occurrents." This way of being and structure of entities in the world can be viewed as reflected in human language. It is a linguistic phenomenon known as the distinction of "individual-level predicates" (ILPs) and "stage-level predicates" (SLPs). In Carlson (1977), ILPs (e.g., *be big*) are defined as *properties attributed directly to the individual*, while SLPs (e.g., *be open*) are described as *properties attributed to the temporal part of an individual* (page. viii). The existence of this distinction as a linguistic phenomenon suggests that humans have a general capability to make this distinction when they describe entities in the world.

This leads us to the following question. On what characteristics of entities do people make the distinction between ILPs and SLPs? To approach this question, we use image-caption data. Image caption data, in which humans describe the contents of (photograph) images, has been provided as training data for machine learning models in recent AI research, especially in the field of Vision and Language (cf. Bernardi et al., 2016; Garcia-Garcia et al., 2018). This type of dataset has several advantages in cognitive science. First, it consists of a large amount of data, in the order of hundreds of thousands. Second, this leads to a diversity of scenes and their corresponding linguistic descriptions. Third, the fixed viewpoint, similar to that of a photograph, allows for easier experimental control compared to field research in the real world. Overall, our approach may shed light on the semantic relationship of the sentence being grounded in the referred world, as depicted in images (Bender & Koller, 2020; Søgaard, 2023; Sato & Mineshima, 2024).

Since we use photograph-images instead of entities-in-the-world, our question can be rephrased as follows: what features of the image are responsible for or contribute to the distinction between ILPs and SLPs? To address the question, the first step we select in Section 2 is to identify the ILPs and SLPs vocabularies that are frequently used in image-caption. Compared to previous corpus studies (e.g., Govindarajan, Durme, & White, 2019), it is distinctive to annotate against image-caption text and to do so for Japanese data. The second step we select in Section 3 is to qualitatively analyze the correspondence between image and text (caption). Qualitative analysis of image caption data has been studied in some depth. In particular, the following studies have been con-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ducted on negation and images: van Miltenburg et al. (2016), Sato et al. (2023), and Berger et al. (2023). See also Sato & Mineshima (2022) for the analysis of universal quantifiers. The third step we select in Section 4 is to build machine learning models and compare their performance to human performance. Here, we use image classification tasks since we aim to know what features of an image contribute to the distinction between ILPs and SLPs by seeing if an AI model trained only on the image data produces the same output results as a human. This method was used by Sato et al. (2023). We adapt it for application to the current issue.

We have given two levels of annotation (manual and automatic) for ILPs and SLPs in this study. In Section 2, a limited amount of text data (around 500 cases) of image captions is manually annotated for ILPs and SLPs by a Japanese linguistics expert. Then, in Section 3, based on the manual annotations, automatic annotation is applied to all data in the dataset (approximately 400,000 items). Section 3 also provides a qualitative analysis of the correspondence between the images and text data. In Section 4, training data is built based on the automatic annotations, and machine learning experiments for image classification are conducted.

Characterizing image-caption sentences as individual-level or stage-level

The SLP/ILP distinction has been widely recognized since it was defined in Milsark (1974) and Carlson (1977), as it is related to various grammatical phenomena across languages. For example, it is a well-known fact that the Spanish *be*-form becomes *estar* in SLP and *ser* in ILP (Diesing 1992). Similarly, in Japanese, the SLP/ILP distinction has received much attention (Masuoka 1987, 2021; Kageyama 2009, 2012; Suzuki 2022, among others), and it is known that the selectivity of the particle *ga/wa* to mark the subject changes depending on whether the predicate is SLP or ILP (Kuno 1973; Heycock 1993, 2008, among others). More specifically, one usage of *ga*, known as descriptive *ga*, does not co-occur with ILPs, though the conflict is not as formally obvious as in Spanish.

Previous studies, such as Milsark (1974), cite predicates for temporary states, such as *sick*, *drunk*, and *tired*, as examples of SLPs, and predicates for permanent states, such as *tall*, *intelligent*, and *beautiful*, as examples of ILPs. Most nouns are considered to be ILPs, while adjectives and verbs can be either SLPs or ILPs. Many non-stative verbs belong to SLP, including examples such as *standing on a chair*, while stative verbs such as *having long arms* are examples of ILP. Fernald (2000) proposes the following descriptive generalization about the correlation between SLP/ILP classification and stativity: All eventive predicates are SLPs, and all ILPs are stative predicates. However, the generalization that all SLPs are eventive predicates does not hold because of the existence of SLP stative predicates such as *being in the room*.

In conducting our corpus study, we hypothesized that stage-level predicates (SLPs) would be more common than individual-level predicates (ILPs) in image caption data. A related study is Alikhani and Stone (2019), which analyzes the distribution of verbs in caption data and shows that captions prefer present progressive. As Diesing (1992) states that the progressive form is an indicator of the stage-level Infl, present progressive forms can be regarded as SLPs in English. Considering the results of Alikhani and Stone's study in light of our study, the tendency that the present progressive is the dominant verb in the captions leads to the prediction that caption data contain more SLPs than ILPs. However, when shifting the focus to Japanese, we cannot equate the English progressive form with the Japanese *-teiru* form. This is because the aspectual meaning expressed by *-teiru* form in Japanese is not limited to progressive but covers a wide range of meanings, including "resultative", "experiential", "iterative/habitual", and "stative" (Kindaichi 1950, Kudo 1995, Kaufmann 2020 among others). Therefore, in order to clarify the distribution of predicates in Japanese caption data, it is necessary to distinguish SLP or ILP by considering the meaning represented by the sentence as a whole, rather than judging only by tense and aspectual form.

Annotation: Method

We randomly selected 530 sentences from STAIR Captions (Yoshikawa, Shigeto, & Takeuchi, 2017), a dataset of Japanese captions. One of the authors, an expert of Japsnese linguistics in this field, annotated them with either SLP or ILP and the final decisions were made in consultation with the other authors. Even though the caption data were expected to consist of relatively simple sentences, most examples were composed of complex sentences. Accordingly, we annotated both the matrix and subordinate clauses separately. For example, in (1), both the subordinate and matrix predicates ("on a skateboard" and "is jumping") are labeled as SLP; in (2), the subordinate predicate ("red") is labeled as ILP and the matrix predicate ("is riding") as SLP; in (3), both the subordinate and matrix predicates ("with the clock tower" and "is made of bricks") are labeled as ILP.[1]

(1) Sukeboo-ni not-ta dansee-ga janpushi-teiru
skateboard-DAT get.on-PST man-NOM jump-TEIRU
'A man on a skateboard is jumping'

(2) Josee-ga aka-i baiku-ni matagat-teiru
woman-NOM red-NON.PST motorcycle-DAT ride-TEIRU
'A woman is riding a red motorcycle'

(3) Tokeedai-ga a-ru tatemono-wa renga-de
clock.tower-NOM have-NON.PST building-TOP brick-INS
deki-teiru
be.made-TEIRU
'The building with the clock tower is made of bricks'

It is clear from the comparison of (1)–(3) that similar predicates in terms of the use of *-teiru* form are used as SLPs in some cases and as ILPs in others. Therefore, instead of using the form alone as a clue for annotation, the meaning of

[1] Abbreviations used in the glosses are the following: NOM nominative; DAT dative; PST past; NON.PST nonpast; TOP topic; INS instrumental; GEN genitive.

the verb, the nature of the subject and even world knowledge should also be taken into consideration. We labeled as ILP those instances that represent permanent or temporally stable states. In contrast, instances that represent events or states deviating from the subject's default state were marked as SLP. While permanent states are always classified as ILP, states that cannot be parmanent, like *atarashi-i* (new) or *furu-i* (old), are annotated as ILP if they represent a relatively temporally stable stages. Although ILPs generally correspond to predicates that characterize subjects, predicates that describe relationships between the properties of multiple individuals, such as *onaji-da* (same) and *chiga-u* (different), are sometimes difficult to determine what they are characterizing. In our study, such predicates are considered ILPs because of their temporal stability. We also consider sentences such as specificational sentences and pseudo-cleft sentences as ILPs, although it is difficult to say that they are characterizing their subjects.

An example of a case where it is difficult to determine whether a predicate is SLP or ILP is the description of clothing and equipment, which often occurs in subordinate clauses. These can be considered SLPs or ILPs depending on what they are wearing and who is wearing them. Clauses such as *a woman wearing a red necklace* are annotated as SLP, while clauses such as *a dog wearing a red collar* are annotated as ILP. One of the criteria for judging such difficult cases is whether or not the clause can express habitual meaning without a quantifying adverb such as *always*. For example, when comparing *a woman (always) wears a red necklace* and *a dog (always) wears a red collar*, the latter can express a habitual meaning without *always*, but the former cannot, so the latter is judged to be an SLP and the former is judged to be an ILP.

We assigned the "ILP/SLP" tag to the few cases that are equally likely to be SLPs and ILPs. In our annotated data, some instances fell outside the scope of the investigation, so we labeled these as "others." For example, nominals such as (4) and meta-references such as (5) are marked as "others." Consequently, sentences without explicit subjects such as (6) are treated as meta-references when supplemented with a subject, as in *The dish in this photo is a stir-fried cauliflower and broccoli*, and uniformly categorized as "others."

(4) shawaa-to basutabu-to senmendai-to yooshikibenki
shower-and bathtub-and washstand-and western.style.toilet
'Shower, bathtub, washstand, and Western-style toilet'

(5) oshare-na tokeedai-ga syashin-ni
fashionable-attributive.form clock.tower-NOM photo-in
utsut-teiru
capture-TEIRU
'A fashionable clock tower is captured in the photo'

(6) Karifurawaa-to burokkorii-no itamemono dea-ru
cauliflower-and broccoli-GEN stir.fry be-NON.PST
'(It is) a stir-fry of cauliflower and broccoli'

Table 1 shows the statistics of each tag in matrix and subordinate clause. Additionally, to identify the characteristics of predicates used in matrix clauses of captions in each case of SLP and ILP, we listed those appearing with high frequency in Table 2. Notably, while verbs like *a-ru* (to be) and *oi-tearu* (to be placed) were also frequently used, they were excluded from the list because both SLP and ILP instances were observed. This is because sometimes an object just happens to be present (or placed) temporarily, while other times it remains there for a longer duration.

Table 1: Number of occurrences of ILPs and SLPs: *matrix* refers to matrix clauses and *sub* refers to subordinate clauses.

ILP		SLP		ILP/SLP	others
matrix	*sub*	*matrix*	*sub*		
52	326	347	324	9	24

Table 2: Top 5 ILPs and SLPs in matrix clauses in frequency. Numbers in parentheses show each predicate's occurrences.

Type	**Example**
ILP	*tui-teiru* (be attached) (5)
	kazarare-teiru, *kazat-tearu* (be displayed) (4)
	de-aru (be) (copula) (3)
	kai-tearu, *kakare-teiru* (be pictured/written) (3)
	settisi-tearu (be installed) (3)
SLP	*(ga) iru* (be located) (23)
	suwat-teiru, *suwa-ru* (be sitting) (13)
	arui-teiru (be walking) (13)
	tat-teiru (be standing) (13)
	tomat-teiru (be parked, be stopped)) (13)

Results and discussion

Table 1 shows that SLPs were significantly more prevalent in the matrix clauses, with fewer occurrences of ILPs. In contrast, in subordinate clauses, it was observed that SLPs and ILPs appeared in roughly equal numbers.

This difference in the distribution of SLPs and ILPs may be due, in part, to the difference in the function of the expressions that identify the objects in the subordinate clauses and those that make predication in the matrix clauses. Typically, the former represents background information, while the latter represents information in focus. When identifying an object, one can use its stable properties to determine what type of object it is. Artifacts and their default states such as *red motorcycle* are typical examples. On the other hand, what is predicated in the matrix clause is typically its motion or temporal state, which is represented by SLPs.

Greenberg's (2011, 2021) pictorial semantics is a suggestive precedent study that may help explain this contrast. According to it, the truth or accuracy of images (pictures) is defined in relation to a viewpoint involving a specific time and space. Recognizing temporal properties in images is challenging because a static image captures only a single moment. By contrast, identifying an object in a particular scene is relatively straightforward to achieve with the constant properties of that object such as color and size. This view aligns with the observation that ILPs tend to appear in subordinate clauses.

Qualitative analysis of image and captions: What images do relate to ILPs and SLPs?

Sentences containing the top five predicates in each of the SLPs/ILPs in Table 2 were extracted from the entire STAIR caption training data (413,915 cases) and analyzed for image-sentence correspondence after morphological analysis and POS tagging using Spacy and ja-GiNZA.[2] STAIR caption (Yoshikawa, Shigeto, & Takeuchi, 2019) is a Japanese caption for Microsoft COCO image data (Lin et al. 2014), with five captions per image (by five human annotators).

In this study, a criterion of more than 3 (out of 5) was employed for determining correspondence between an image and a sentence. If this threshold was exceeded, the pair was considered to have a correspondence. However, it is important to note that it is not always feasible to determine whether a predicate is stage-level or individual-level solely based on the predicate itself. Therefore, we will conduct a qualitative analysis of the examples below.

Images and sentences for individual-level information (ILPs)

For sentences containing the ILPs *be attached* ("tui-teiru"), 27 images surpassed the criterion. Figure 1 lists 3 images that passed the even higher criterion of 4/5 or more. For example, the following captions (translated into English) are given for the image of attach(a):

1. *Lights of various sizes are attached to the bike.*
2. *Lots of lights are attached to the bike.*
3. *Lots of headlights attached to a white car.*
4. *Many round lights are attached to the front of a mod bike.*
5. *Lights are attached to the front of a vehicle.*

The third was excluded because it has no subject and is not a sentence. The other four captions contain sentences with the ILPs related to the predicate *be attached*. These seem to possess the characteristics of ILPs, specifically being independent of a specific time and not temporary.

The same is true for the other two cases attach(b) and attach(c) images in Figure 1. The two cases correspond to the sentences *a chain is attached* and *a signboard is attached*, respectively. In Japanese, the verb "tui-teiru" is ambiguous between an ILP interpretation and an SLP interpretation. Thus, examples such as *the light is on* ("akari-ga tui-teiru"), *trash is attached* ("gomi-ga tui-teiru"), and *resting his elbow* ("hizi-o tui-teiru") were not classified as ILPs.

For sentences containing the ILPs *be displayed* ("kazarare-teiru"), 85 images surpassed the criterion. Figure 1 lists 14 images (display(a–n)) that passed the even higher criterion of 4/5 or more. Sentences like *flowers and/or paintings are displayed* corresponds to these images.

For sentences containing the copula *be* ("de-aru") which is an ILP, only one image surpassed the criterion, shown in be(a) of Figure 1. Sentences such as *the walls is black and white border pattern* correspond to this image. Another example that exceeded the criterion was the case described as *the traffic signal is red* (COCO id #178168), but this was not counted as ILPs because of the characteristic of traffic signals that change color within a short period of time.

For sentences containing the ILPs *be written* ("kakare-teiru"), 71 images surpassed the criterion. Figure 1 lists 12 images (write(a-l)) that exceeded an even higher criterion of 4/5 or more. These images correspond to sentences such as *pictures and/or words are written*. The instance with the active voice, *he is painting something*, is not classified as ILPs.

For sentences containing the ILPs *be installed* ("setti-site-aru"), no image met the criterion.

Images and sentences for stage-level information (SLPs)

In all individual-level descriptions, the subject was inanimate. On the other hand, there were few cases where the subject was inanimate in the stage-level descriptions. As described in Section 2, the tendency for SLPs images to be more common than ILPs images in the image caption data is also true for the entire STAIR caption training data. For simplicity's sake, let us use the 5/5 criterion. According to this criterion, 260 images with SLPs surpassed it (14 for *be located* ("iru"), 21 for *be standing* ("tat-teiru"), 47 for *be sitting* ("suwat-teiru"), 104 for *be stopped* ("tomat-teiru")). Meanwhile, 5 images with ILPs met the criterion (2 for *be displayed* ("kazarare-teiru") and 3 for *be written* ("kakare-teiru")). Nevertheless, of the former 260, only 49 of *be stopped* (as in *the car is stopped/parked*) had an inanimate object as the subject, while the other 211 had an animate target as the subject.

The upper row of Figure 2 shows a case in which a sentence containing a SLPs image passed the 5/5 criterion and had an animate subject. The images include: *be located* (*elephants are located*), *be sitting* (*a man is sitting on a bench*), *be walking* (*elephants are walking*), *be standing* (*a giraffe is standing*), *be stopped* (*a bird is perched on a branch*), *be riding* (*a man and a dog are riding on a bike*), and so on.

As for *standing*, there were some cases like *the sign stands up*; four cases were identified in the 5/5 criterion. However, these were rather ILPs cases and were not counted as SLPs cases.

When the 4/5 criterion was used, there were also cases of inanimate objects in *sitting* (Figure 2). Although *sit* is used only for humans and animals, there were cases where it was used for stuffed animals and dolls, which were also counted as SLPs, albeit debatable.

However, since this was not sufficient to include inanimate objects, we also analyzed *be riding*, which was next in the top five. 21 images met the 5/5 criteria, all of which had an inanimate object as the subject. But when the criteria were reduced to 4/5, there were 13 cases of inanimate objects as the subject (*a dish is on a plate*) (Figure 2). Note that both *be riding* and *be on* in English are translated into *not-teiru* in Japanese.

[2] https://github.com/megagonlabs/ginza

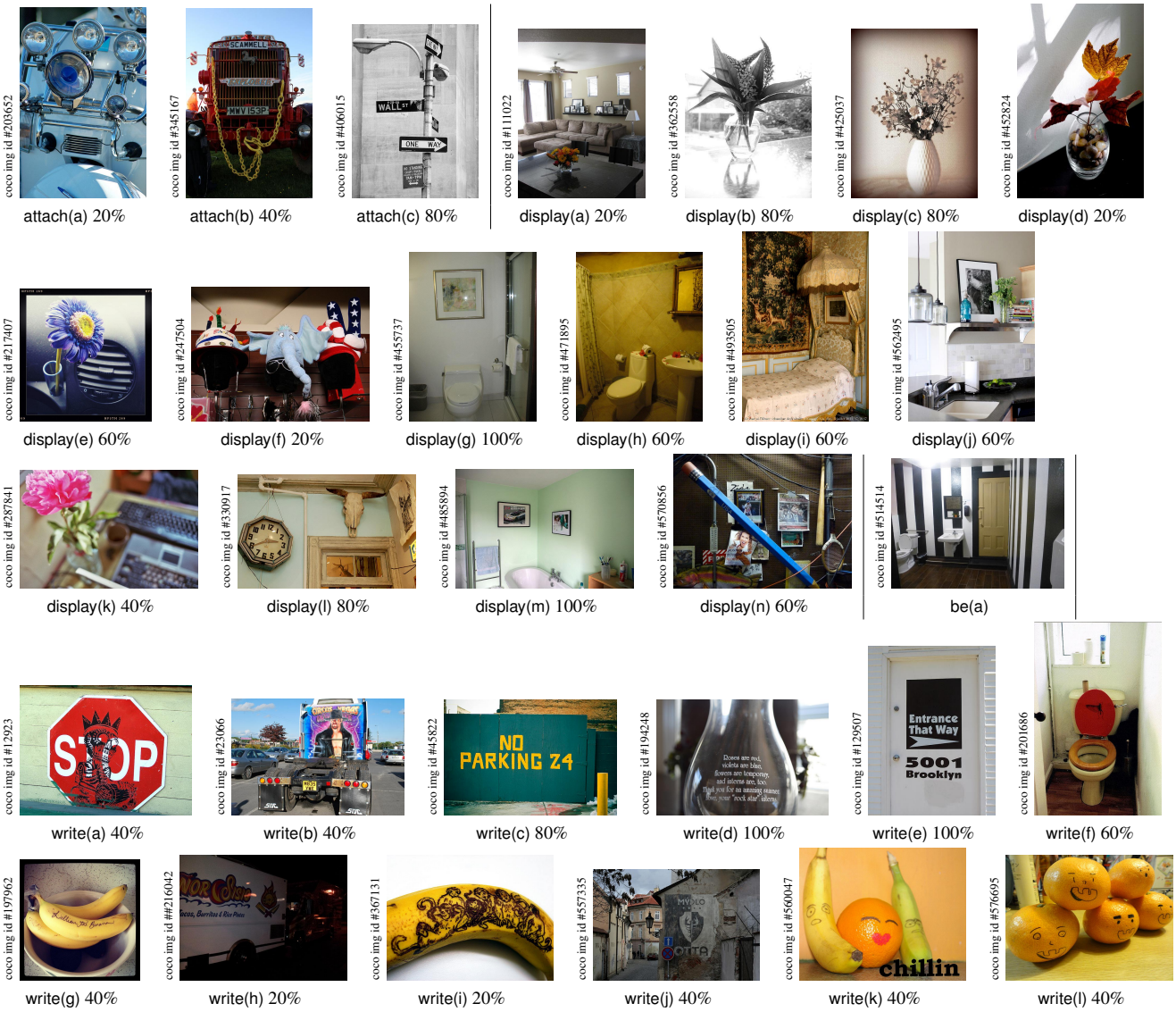

Figure 1: Images describing scenes at an individual level (ILPs): all items other than be(a) meet the 4/5 criteria and percentages mean models' average accuracy in the classification tasks

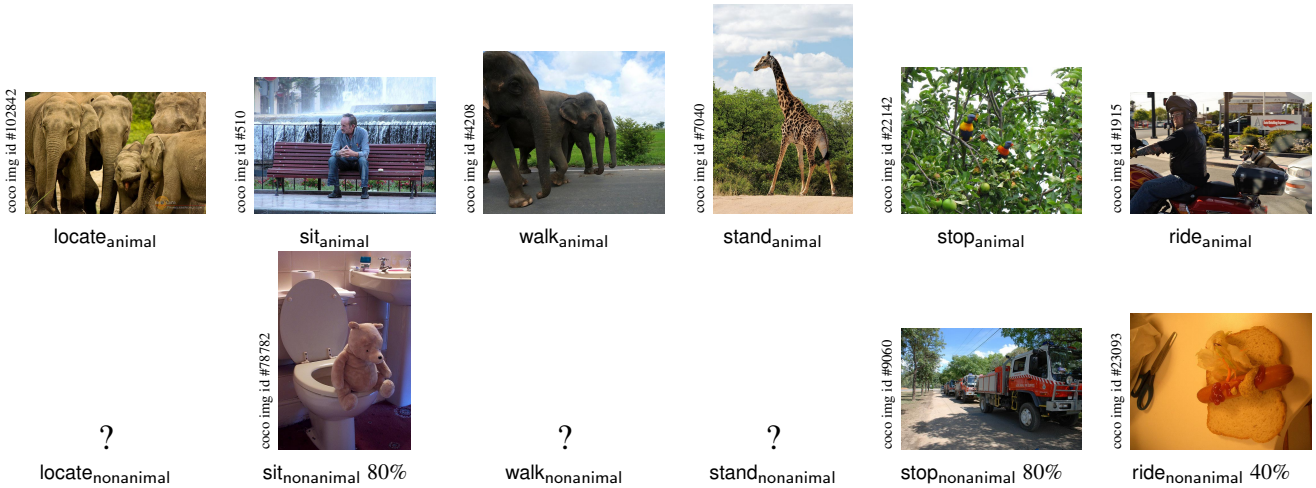

Figure 2: Images describing scenes at a stage level (SLPs); "?" means that there is no corresponding image and percentage mean models' average accuracy in the classification tasks

Can machine learning models classify images as expressing ILPs and SLPs?

In this section, we build machine learning models and compare their performance to human performance. Here, we use image classification tasks since we aim to know what features of an image contribute to the distinction between ILPs and SLPs by seeing if an AI model trained only on the image data produces the same output results as a human.

Method

There were 183 ILPs images that met 3/5 or better, and 29 of all images that matched at 4/5 of Figure 1 were used as the test images. The remaining 154 images (images matching 3/5) were used as a set of train and validation images: 24 for attach, 71 for display, and 59 for write.

183 images were also selected for SLPs. However, since it was necessary to minimize and control for factors other than ILPs/SLPs (differences in images), we used only inanimate subjects for SLPs. The 29 test images included all 4 images that met the 4/5 criteria for sit, while the remaining consisted of 12 for ride and 13 for stop. For ride, out of the 13 images that matched at 4/5, 12 were selected. For stop, 13 out of the 49 images that matched at 5/5 were chosen in ascending order of their image ID numbers. For the set of 154 train and validation images, sit matched all 11 images at 3/5, while ride and stop were selected from images that matched at 3/5 or above, totaling 71 and 72 images respectively.

For the set of train and validation images, image augmentation was performed with horizontal, vertical, and both flips. 616 images (4 times of 154) were prepared for each of ILPs and SLPs. The file numbers were randomized and 500 images were used for training and 116 for validation. This procedure was repeated 5 times, and each set of images was used in models 1–5. In other words, the images and the order in which they were used for training were set up differently among the models.

A convolutional neural network (CNN) model with the VGG16 fine-tuning (Simoyan & Zisserman, 2015) was used. We built models in Python's Keras with these settings: sequential model, intermediate layers using ReLU, output layer using Softmax, 0.5 dropout rate, first 14 layers' weights from VGG16, Cross-Entropy loss, batch size of 18, and 3 epochs.

Results

Table 3 shows the results of the five CNN model tests. The accuracy rates for ILPs were 62.1%, 55.2%, 48.3%, 37.9%, 72.4% (mean 55.2%) and for SLPs were 55.2%, 79.3%, 86.2%, 82.8%, 75.9% (mean 75.9%). 4 out of 5 (80% of) people used ILPs in the ILPs images in the test. 4 out of 5 people used SLPs in the 16 items of SLPs images and 5 out of 5 people used SLPs in the 13 items of SLPs images (thus 88.9%). The findings suggest that the model's performance of ILPs is about chance level, while its accuracy on SLPs is lower than human performance but still reasonably correct.

Table 3: Accuracy results of image classification task for machine-learning (CNN+VGG16) models

	model 1	model 2	model 3	model 4	model 5	average
ILPs	62.1%	55.2%	48.3%	37.9%	72.4%	55.2%
SLPs	55.2%	79.3%	86.2%	82.8%	75.9%	75.9%

General discussion

A limited amount of Japanese image caption data were randomly extracted and given annotation with SLPs/ILPs. The results showed that the majority of the data consisted of SLPs, with a small proportion being ILPs. We identified predicates with high frequency of occurrence. Based on manual annotation, we conducted automatic annotation for the entire dataset and machine learning experiments for image classification based on ILPs and SLPs. Our results showed that the accuracy rate for SLP judgments was significantly high, though not as high as human performance, whereas the accuracy rate for ILP judgments was around chance level, lower than human performance. Since ILPs were the primary focus of this experiment, we restricted SLPs to inanimate subjects to align with the ILPs' inanimate subjects. This approach may have skewed the natural diversity of SLPs, which typically include many animate subjects. Consequently, the high accuracy in SLP judgments could be attributed to the limitation imposed on the inherently varied SLP subjects to a narrow range, such as cars and food. The low accuracy to images in which cars and fruits appear in the ILPs may be the result of the bias that limited these subjects to inanimate subjects of the SLPs.

The ILP judgments underperformed, but what is the reason behind this? One potential factor is the diversity in the subjects or targets. First, within the predicates selected during the selection process, there is a wide range of subjects. For instance, display has subjects such as vases, paintings, and other objects, with even greater diversity observed in training data due to the difference in criteria for test (4/5) and training images (3/5). Additionally, there are other predicate categories that exhibit similar diversity (a total of three in this experiment), and ILPs are identified as a common characteristic among them. In this respect, ILPs judgments require handling higher-order categorization judgments, which can cause difficulties. This might also relate to "indirect grounding" of abstract concepts (Cerini, Di Palma & Lenci, 2022; Utsumi, 2022), highlighting a challenge for future AI research.

A problem with the photograph caption data used is that ILPs data is extremely scarce compared to SLPs data. One approach here, as suggested in the context of way-finding scenarios in the introduction, is to incorporate a "purpose" within the AI system. Another approach is to use as training data images that reflect situations in which the use of ILPs information is naturally required. Images that visually represent time-independent information, such as diagrams as used in a mathematical context (Allwein & Barwise, 1995) or illustrations as used in a natural science context (Kembhavi et al, 2016), could be effective for this. The pursuit of these ideas remains an interesting future challenge.

Acknowledgments

This study was supported by Grant-in-Aid for JSPS KAKENHI Grant Number JP24K15676, JP24K00004, JP19K13172, as well as JST CREST Grant Number JPMJCR2114.

References

Alikhani, M., & Stone, M. (2019). "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language* (pp. 58–67). ACL.

Allwein, G. & Barwise, J., Eds. (1995). *Logical reasoning with diagrams*. New York: Oxford University Press.

Arp, R., Smith, B., & Spear, A. D. (2015). *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press.

Bender, E. M. & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185—5198), ACL.

Berger, U., Frermann, L., Stanovsky, G., & Abend, O. (2023). A Large-Scale Multilingual Study of Visual Constraints on Linguistic Selection of Descriptions. In *Proccedings of 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2240–2254), ACL.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55, 409–442.

Carlson, G. (1977). *Reference to Kinds in English*, PhD thesis, University of Massachusetts.

Cerini, L., Di Palma, E., & Lenci, A. (2022). From Speed to Car and Back: An Exploratory Study about Associations between Abstract Nouns and Images. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment* (pp. 80–88). ACL.

Diesing, M. (1992). *Indefinites*. Cambridge, MA: MIT Press.

Fernald, T. B. (2000). *Predicates and Temporal Arguments*. Oxford: Oxford University Press.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., & Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70, 41–65.

Govindarajan, V., Durme, B. V., & White, A. S. (2019). Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7, 501–517.

Greenberg, G. (2011). *The Semiotic Spectrum*. PhD thesis. Rutgers University.

Greenberg, G. (2021). Semantics of pictorial space. *Review of Philosophy and Psychology*, 12(4), 847–887.

Heycock, C. (1993). Focus projection in Japanese. In Mercè, G. (ed.) *Proceedings North East Linguistic Society, Vol. 24*, (pp.159–187), Amherst, MA: GLSA Publications.

Heycock, C. (2008). Japanese -wa, -ga, and information structure. In Miyagawa, S. & Saito, M. (eds.) *The Oxford Handbook of Japanese Linguistics* (pp.54–83), Oxford University Press.

Kageyama, T. (2009). Gengo no Koozoo seiyaku to jojutsu kinoo [Structural constraints and predication functions in language], *Gengo kenkyuu [Linguistic studies]* 136 (pp.1–34).

Kageyama, T. (2012). Zokuseejojutsubun no bunpooteki igi [Grammatical significance of property predication], In Kageyama, T. (ed.) *Zokuseejojutsu no sekai [World of property predication]* (pp.3–35). Tokyo: Kurosio.

Kaufmann, S. (2020). Formal treatments of tense and aspect. In Wesley M. J. & Takubo Y. (eds.) *Handbook of Japanese Semantics and Pragmatics* (pp.371–422). Berlin: De Gruyter Mouton.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. In *Proceedings of the 14th European Conference on Computer Vision, LNCS vol 9908* (pp. 235–251), Springer.

Kindaichi, H. (1950). Kokugo dooshi no ichi bunrui [A classification of Japanese verbs], *Gengo kenkyuu [Linguistic studies]* 15, pp.48–56.

Kudo, M. (1995). *Asupekuto, tense taikei to tekusuto: Gendai nihongo no jikan hyoogen [Aspect, tense system, and text: The expression of time in contemporary Japanese]*. Tokyo: Hituzi.

Kuno, S. (1973). *The Structure of the Japanese Language*. Cambridge, MA: MIT Press.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., & Perona, P. (2014). Microsoft COCO: Common objects in context. In *Proceedings 13th European Conference on Computer Vision*, LNCS vol 8693 (pp. 740–755), Springer.

Masuoka, T. (1987). *Meidai no bunpoo: Nihongo bunpoo josetsu [Grammar of propositions: Introduction of Japanese grammar]*. Tokyo: Kurosio.

Masuoka, T. (2021). *Nihongo bunron yookoo: Jojutsu no ruikee no kanten kara [Theory of Japanese sentences: From the perspective of description types]*. Tokyo: Kurosio.

Milsark, G. L. (1974). *Existential Sentences in English*. PhD thesis, MIT.

van Miltenburg, E., Morante, R., & Elliott, D. (2016). Pragmatic factors in image description: the case of negations. In *Proceedings of the 5th Workshop on Vision and Language* (pp. 54–59). ACL.

Sato, Y., & Mineshima, K. (2022). Visually analyzing universal quantifiers in photograph captions. In *Proceedings of 13th International Conference on the Theory and Application of Diagrams*, LNAI vol. 13462, (pp. 373–377), Springer.

Sato, Y., Mineshima, K., & Ueda, K. (2023). Can Negation Be Depicted? Comparing Human and Machine Understanding of Visual Representations. *Cognitive Science*,

47(3), e13258.
Sato, Y. & Mineshima, K. (2024). Can machines and humans use negation when describing images? In *Proceedings of 2nd International Conference on Human and Artificial Rationalities*, LNCS vol. 14522 (pp.39–47), Springer.
Simoyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Paper presented at *the International Conference on Learning Representations, 2015, San Diego, CA, May 7–9, 2015.*
Suzuki, A. (2022). *Zokuseejojutsu to soosyoosee [Property predication and genericity]*. Tokyo: Kachoosya.
Søgaard, A. (2023). Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33, 33–54.
Utsumi, A. (2022). A test of indirect grounding of abstract concepts using multimodal distributional semantics. *Frontiers in Psychology*, 13, 906181.
Yoshikawa, Y., Shigeto, Y., & Takeuchi, A. (2017). STAIR captions: Constructing a large-scale Japanese image caption dataset. In *55th Annual Meeting of the Association for Computational Linguistics* (pp. 417–421). ACL.
Zemach, E. (1970). Four Ontologies. *Journal of Philosophy*, 67(8), 231–247.