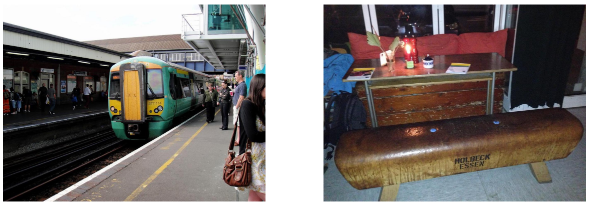
Describing Images *Fast and Slow*: Quantifying and Predicting the Variation in Human Signals during Visuo-Linguistic Processes

Ece Takmaz and **Sandro Pezzelle** and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{ece.takmaz|s.pezzelle|raquel.fernandez}@uva.nl

Abstract

There is an intricate relation between the properties of an image and how humans behave while describing the image. This behavior shows ample variation, as manifested in human signals such as eye movements and when humans start to describe the image. Despite the value of such signals of visuo-linguistic variation, they are virtually disregarded in the training of current pretrained models, which motivates further investigation. Using a corpus of Dutch image descriptions with concurrently collected eye-tracking data, we explore the nature of the variation in visuo-linguistic signals, and find that they correlate with each other. Given this result, we hypothesize that variation stems partly from the properties of the images, and explore whether image representations encoded by pretrained vision encoders can capture such variation. Our results indicate that pretrained models do so to a weak-to-moderate degree, suggesting that the models lack biases about what makes a stimulus complex for humans and what leads to variations in human outputs.



Min: 1.69 sec   Max: 7.07 sec

Figure 1: The images with the minimum and maximum mean speech onsets across speakers in the dataset. The image with the maximum onset also elicits the highest variation in the first nouns of the descriptions.

1 Introduction

Humans can capture the gist of an image usually incredibly fast – 100 msec could be enough (Oliva, 2005; Oliva and Torralba, 2006); however, they would need more time to act on an image. For instance, human behavior while describing images illustrates the intricacies of visuo-linguistic processes. There may be repetitions, silent intervals and disfluencies, with considerable degrees of variation in what is uttered across speakers. The period prior to the utterance involves perceiving the image, conceptualizing the message, retrieving the labels of the entities to mention, formulating and preparing to articulate a grammatical and relevant utterance (Levelt, 1981; Slobin, 2003).

As a result, we observe variations in **speech onsets**, as in Figure 1, which could be indicative of the relative cognitive complexity induced by the images (Coco and Keller, 2015; Gatt et al., 2017). In addition, different speakers might start their utterances with different words (**starting points**, see MacWhinney, 1977), continuing to produce a varied set of image descriptions (**linguistic variation**) with **variation in gaze**. These signify the intricate cross-modal relation between visual and linguistic processes in humans (Griffin and Bock, 2000; Ferreira and Rehrig, 2019).

Although human data can be rich in behavioral signals, current pretrained multimodal models virtually never receive information about such signals during training. The models generate descriptions without necessarily modeling how human processes unfold. For instance, deep neural networks can output words at the same rate even for images that would result in diverse speech behavior by humans due to complexity or ambiguity. Moreover, there is a gap between the manner in which humans perceive stimuli as compared to how large models process them. Model-predicted surprisal values for linguistic input can be lower than human surprisal, possibly due to the massive size of the training data and the number of model parameters (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Oh and Schuler, 2023a,b). Models also display different patterns of visual attention

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2072–2087

compared to humans (Das et al., 2016).

We argue that it is essential to consider human signals such as speech onsets and looking times, as they reflect the complexity and ambiguity of visuo-linguistic tasks (Coco and Keller, 2015; Gatt et al., 2017; van der Meulen et al., 2001; Meyer and van der Meulen, 2000; van Miltenburg et al., 2018b). It is therefore desirable if models encode what leads to variations in such signals to help generate image descriptions in a way that is aligned with human processing and with types of variation observed in human data (van Miltenburg et al., 2018a). To this end, several applications have exploited human gaze to enhance image captioning and visual question answering models (Sugano and Bulling, 2016; He et al., 2019; Takmaz et al., 2020; Sood et al., 2021, 2023). Still, the relation between gaze on images and language is not widely researched in NLP (Alacam et al., 2022).

We first explore the natural dynamics in visuo-linguistic processes using The Dutch Image Description and Eye-tracking Corpus (DIDEC; van Miltenburg et al., 2018b). This corpus provides gaze and speech data concurrently collected while participants describe images depicting real-life scenes. We preprocess the DIDEC dataset extensively, and propose metrics to quantify the variation in visual and linguistic modalities. We reveal for the first time significant correlations between speech onsets, variation in starting points, descriptions and gaze.

We hypothesize that this variation is partly due to the properties of the images, and that similar images would elicit similar amounts of variation. Given the superior performance of pretrained encoders that are widely used in multimodal models, we investigate whether visual encoders such as CLIP (Radford et al., 2021) and ViT (Dosovitskiy et al., 2021) capture information regarding the variation in visuo-linguistic signals.[1] This is akin to probing pretrained models for meaningful syntactic and semantic information; see Conneau et al., 2018. Using a similarity-based prediction method (Anderson et al., 2016), we find that the pretrained encoders capture variation in signals to a limited extent. Our findings suggest that underlying factors leading to variation are encoded rather weakly by pretrained models. With our work, we aim to direct attention towards the importance of the information contained in such signals and the variation thereof when crowdsourcing data as well as during model development.

[1] Code and data available at `https://github.com/ecekt/visuolinguistic_signal_variation`.

2 Background

We first give an overview of visuo-linguistic processes in humans in Section 2.1, and then, in models in Section 2.2.

2.1 Visuo-Linguistic Processes in Humans

Cross-modal processes Describing images requires the linear unfolding of complex cross-modal processes between vision and language (Henderson and Ferreira, 2013; Griffin and Bock, 2000; Gleitman et al., 2007; Coco and Keller, 2012; Ferreira and Rehrig, 2019; Henderson, 2017). There exist several theories regarding how the 'linearization' (Levelt, 1981) takes place in sentence formulation in relation to visual processes (Griffin, 2004; Meyer, 2004; Ferreira and Rehrig, 2019). These theories consider the speaker's knowledge and expectation regarding the contents of the image, as factors affecting the allocation of gaze and the formulation of a description (Henderson, 2017; Ferreira and Rehrig, 2019). In addition, the way people look at an image changes based on the task at hand (Yarbus, 1967; Buswell, 1935; Castelhano et al., 2009), with similar sequences of fixations (*scanpaths*) leading to the production of similar sentences (Coco and Keller, 2012). Therefore, we hypothesize that the variation in language production and eye movements could be correlated.

Starting points A sentence must have a starting point, given that words need to be uttered in a linear order (Levelt, 1981). We take the first uttered noun as the starting point of image descriptions. The focus on nouns is motivated by the fact that gaze scanpaths are frequently represented by sequences of object categories, which tend to be expressed by nouns. Additionally, the order of mention of these categories is the point of interest in linearization studies that investigate language production parallel to visual processes (Ferreira and Rehrig, 2019). Starting points can be selected based on a variety of factors (canonical word order of the language, perspective of the speaker, complexity of the planned sentence; see MacWhinney, 1977). When describing images, visual properties of an image influence how a sentence begins and unfolds (Bock et al., 2004). These findings signify how the selection of starting points can be influenced by a set of complex visuo-linguistic factors.

Variation in image descriptions People generally describe images with some variation. Jas and Parikh (2015) report that images with people and large objects tend to be described more specifically, whereas generic buildings, ambiguous scenes and images with less-important objects tend to elicit more varied descriptions. The degree to which the descriptions of an image vary is referred to as 'image specificity' by Jas and Parikh (2015), who propose an automatic metric to quantify it using the similarity scores between the WordNet paths of words in descriptions (Miller, 1994). van Miltenburg et al. (2018b) explore image specificity in the corpus that we use in this study, utilizing word2vec vectors (Mikolov et al., 2013) to compute the similarity scores. They find that the variation in descriptions is only to a limited extent due to the image's contents as there also seems to be an effect of language (English vs. Dutch). Additionally, their results indicate that attention maps extracted using gaze data do not help predict image specificity (van Miltenburg et al., 2018b). In this work, we also quantify and predict image specificity proposing different approaches.

Speech onsets Slower speech onsets indicate that a deliberate, effortful process is taking place, as compared to fast onsets; as claimed in the dual process theory (Wason and Evans, 1974; Kahneman, 2012). Various intertwined linguistic and visual processes modulate speech onsets and the latency of referring to an object (Meyer and van der Meulen, 2000; Coco and Keller, 2015), such as the contents of an image and the locations of the objects (Gatt et al., 2017; Esaulova et al., 2019). This indicates that speech onsets are strongly linked to image features. Given the importance of speech onsets in relation to visuo-linguistic processes and the cognitive requirements of a task, the mean speech onset induced by an image across speakers is one of the signals we focus on.

2.2 Multimodal NLP

Pretrained models Many recent multimodal models employ frozen pretrained unimodal models and combine them with either no further training or via trained lightweight mapping networks (Berrios et al., 2023; Alayrac et al., 2022; Mañas et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023; Mokady et al., 2021; Chen et al., 2022). Particularly, the visual encoder of the CLIP model (Radford et al., 2021) has been utilized in these models as a foundation model with strong zero-shot capabilities that improves multimodal models (Shen et al., 2022).

By training classifiers on top of visual encoders, Berger et al. (2023) predict the existence of linguistic features such as passive voice and the use of numeral expressions in image descriptions, and indicate that the selection of such linguistic features is constrained by visual features. These findings point to the underlying capabilities of pretrained models pertaining to human cognitive processes.

Human signals in NLP Most previous research into the use of human signals focuses on text-only cases (Klerke et al., 2016; Barrett et al., 2018, 2016; Mishra and Bhattacharyya, 2018; Hollenstein et al., 2021a, 2022, 2021b; Pouw et al., 2023; Ding et al., 2022; Ren and Xiong, 2021; Dong et al., 2022; Khurana et al., 2023; Mathias et al., 2020; Zhang et al., 2020). However, the relationship between human gaze on images and language production, and its potential contribution to computer vision and NLP has been investigated even before the existence of pretrained models (Yun et al., 2013). Research into whether the attention distributions in multimodal models correlate with human attention reveals contrasting findings (Das et al., 2016; Gella and Keller, 2018; He et al., 2019; Sood et al., 2021). Several works show that the use of human gaze enhances image captioning and visual question answering (Sugano and Bulling, 2016; He et al., 2019; Takmaz et al., 2020; Sood et al., 2021, 2023). Yet, modeling gaze in conjunction with linguistic processes is still an under-explored area in NLP (Alacam et al., 2022).

In our work, we investigate the variation of a set of human signals in a corpus, as well as whether pretrained vision encoders can encode information related to these signals. Although such models are shown to be very effective in multimodal tasks, they are still under-explored from this point of view.

3 Data

We aim to explore the variation in human signals in visuo-linguistic processes and whether pretrained models can capture such variation in a realistic setup. A dataset consisting of simultaneous language production and eye movements over complex images would enable such an exploration. Therefore, we opt for using the DIDEC corpus (van Miltenburg et al., 2018b) instead of other existing image description datasets with eye-tracking, as this corpus allows us to delve into the dynamics of visual and linguistic processes in parallel. There exist few datasets containing such information, which we did not opt for utilizing, as they differ in their tasks (narratives; Vaidyanathan et al., 2018), or the processing steps the authors have taken, e.g., only a small subset of the captions were checked manually (Vaidyanathan et al., 2018); the authors sample one gaze point every 4 points (He et al., 2019). The DIDEC dataset comprises manually checked descriptions of high quality, and the gaze data is provided in a raw format enabling custom processing.

We use the 'production viewing' subset of DIDEC, which contains spoken descriptions for 307 real-life images originating from the MS COCO dataset (Lin et al., 2014), with high-quality eye-tracking data.[2] 45 participants describe ≈ 102 images without a time limit. On average, each image has 15 descriptions (4604 in total). Next, we explain how we extract features corresponding to human signals in visuo-linguistic processes from this dataset, to obtain 4586 descriptions with speech onsets, starting points, and fixated regions.

3.1 Visual Data

Using the raw gaze samples in DIDEC (van Miltenburg et al., 2018b) labeled as fixations, saccades, and blinks, we create fixation windows by treating saccades and blinks as boundaries (Salvucci and Goldberg, 2000). The gaze samples in the fixation window are then put into a list, skipping the ones that fall outside the boundaries of the images. To visually represent a fixation, we feed its gaze points as coordinate prompts to the Segment Anything Model (SAM; Kirillov et al., 2023). Using the prompts, this model predicts the objects the gaze corresponds to, and outputs masks corresponding to fixated regions. We use the ViT-L version of the model building on vision transformers (Dosovitskiy et al., 2021), as it achieves good performance (Kirillov et al., 2023). We obtain a single mask per fixation window. The masks sometimes span non-contiguous regions; therefore, we utilize the bounding box based on the x-y limits of the predicted mask.

3.2 Linguistic Data

Speech onsets The dataset supplies audio files for spoken descriptions and their transcripts. To extract word-level timestamps, we use WhisperX (Bain et al., 2023) based on Whisper (Radford et al., 2023).[3] We relay the transcripts directly into the alignment function of WhisperX. The output contains the start and end timestamps of each word. This also allows us to extract information regarding when the participants start talking, i.e., speech onsets. The mean speech onset is 3.42 sec, and the median is 2.65 sec. We observe variation across participants and images, as the onsets can go up to 25.37 sec with a standard deviation (SD) of 2.45.

Starting points We use the spaCy library for tokenization, part-of-speech tagging, and lemmatization of the words in the descriptions.[4] For Dutch, the library provides 3 models (small, medium, and large). Upon manual inspection of 50 random samples from the data processed by each model, we opted for the large model, which yields the least number of errors. See Appendix A for more details.

4 Variation in Human Signals

We first delve into the nature of the variation across humans per image in the DIDEC dataset. Our focus is on uncovering potential correlations between the variations in human signals in visuo-linguistic processes. We first explain how we quantify each signal and its variation, see Figure 2 for an example image with all of its variation scores. Then, we conduct pairwise correlation analyses between the 4 variables. If there exist correlations between variations across signals, one can speculate that at least part of the correlation stems from the image, with the rest being potentially due to factors such as viewing order, priming and cognitive load.

4.1 Variation in Speech Onsets

We inspect the mean and SD of speech onsets per image, see histograms in Appendix B. The mean onsets per image range between 1.69 and 7.07 seconds, constituting a non-normal distribution skewed towards shorter onsets ($p < .001$, 65.77% of the onsets shorter than the mean onset). For some images, some participants start talking immediately; whereas, in other cases, they wait for a considerable amount of time before speaking. This observation resonates with the fast and slow systems from the dual process theory (Wason and Evans, 1974; Kahneman, 2012), suggesting

[2]The other subset contains data from 'free viewing', where the participants simply looked at the images for 3 seconds.

[3]The model for obtaining alignments for audio in Dutch: `jonatasgrosman/wav2vec2-large-xlsr-53-dutch`

[4]`nl_core_news_lg` pipeline from https://spacy.io/

Mean onset: 3.46 seconds
Variation in starting points: 11
Most common starting point: *pier*
Image specificity BLEU-2: 0.39
Variation in gaze: 38.47

*een **pier** waar het heel erg druk is uh rechts is een vis aquarium waar je vissen kan aanraken*
*(a **pier** where it is very busy uh on the right is a fish aquarium where you can touch fish)*

*een drukke **straat** met een aantal restaurants pier 39*
*(a busy **street** with a number of restaurants pier 39)*

***pier** waar veel mensen lopen*
*(**pier** where many people walk)*

*een drukbezette **pier***
*(a busy **pier**)*

*een toeristische **plaats** waar veel verschillende entertainment dingen te doen zijn*
*(a touristic **place** where there are many different entertainment things to do)*

*de **ingang** van een aquarium met veel mensen op een plein*
*(the **entrance** to an aquarium with many people in a square)*

Figure 2: An image with its variation scores, a subset of its descriptions (along with the English translations in parentheses), and the eye movements of a single participant. In the descriptions, the words in boldface indicate the starting points in Dutch and their equivalents in English.

that more complex processes are recruited while describing certain images. However, even for a single image, the participants might start speaking at varying times (with SD per image ranging from 0.44 to 6.33). This suggests that various factors are at play while describing images, such as contextual and speaker-specific effects.

To have a better picture of onset variation, we compare the onsets for an image against each other. Leaving one onset out of the set of onsets for an image, we calculate the average of the rest (≈ 14 onsets). The difference between the average and the left-out onset corresponds to error. We perform this calculation for each sample. Then, we take the mean over all the samples, which yields an error of 1.625 seconds. This error is a proxy for the average variation over the participants, which suggests that there is a difference in response times across humans when prompted with the same image.

The DIDEC corpus comes with 3 mutually-exclusive image subsets called 'lists'. Each participant views only one list. We find that the mean onsets in List 2 are significantly shorter than the other two sets ($p < .001$, independent samples t-test). Since both the images and the participants are different across lists, it is not straightforward to separate their effects. See Appendix C for a participant-based analysis of mean onsets.

4.2 Variation in Starting Points

Counting the first nouns of image descriptions reveals that there is an imbalance in the starting points in the data. The participants utter words such as *man*, *person*, *woman*, *bus* and *street* most frequently as the first noun of a description (370, 238, 221, 174, 141, respectively, constituting in total 25% of the samples). This is potentially due to the salience of such entities and their frequency in the images. We represent the variation in starting points by the number of unique starting points uttered per image, yielding $mean = 6.45$, $min = 1$, $max = 13$. These values indicate that some images elicit the same first nouns, whereas some others prompt the production of a range of starting points.[5]

4.3 Variation in Full Descriptions

Each image can be described in distinct ways, both in terms of the words uttered and their order. We quantify the *linguistic variation* in image descriptions, following a different approach compared to Jas and Parikh (2015) and van Miltenburg et al. (2018b). We adopt a widely used natural language generation metric, BLEU (Papineni et al., 2002). This metric computes n-gram-based precision scores between a generated sentence and a set of references. We opt for the bigram version (BLEU-2), since we are mostly interested in the surface form variation of words, and to a limited extent, the sequences of words. BLEU-2 allows us to measure the linguistic variation in descriptions independently of a pretrained model.[6] We calculate the BLEU-2 score between a description and the

[5] We compute variation in starting points with respect to exact noun lemma matches, without considering synonyms that could refer to the same object. We believe that this captures the type of variation that is of interest for starting points, since lexical choices reflect categorization and conceptualization of objects that can be affected by the visual context in which the object is situated (Gualdoni et al., 2023).

[6] See Appendices D and E for a semantic variation metric we propose using Dutch BERT-based representations (BERTje; de Vries et al., 2019), another combining BERTje and BLEU-2-based variation, as well as a comparison to human annotations provided by Jas and Parikh (2015).

remaining descriptions for the image constituting the reference set. Then, we take the average over all descriptions of an image.[7] This method yields an extensive range of normally distributed scores ($\mu = 0.53, min = 0.25, max = 0.81$).

4.4 Variation in Gaze

The variation in eye movements has been quantified in various ways in the literature: scanpath complexity, dispersion of the heatmap of gaze on an image, entropy of the gaze distribution (Coco and Keller, 2015). We propose a distance metric based around the contents of fixated regions and their orders. We represent a scanpath in the form of a sequence of fixation bounding boxes represented as (x_1, y_1, x_2, y_2). Given two scanpaths S_1 and S_2, for each fixation box in S_1, we find the most similar box in S_2 that yields the highest ratio of intersection over union (IoU) between the bounding boxes. The IoU dissimilarity $(1 - IoU)$ as well as the normalized positional distance between these boxes are summed up. This step is performed for all fixation boxes in S_1. The total gives us a comparison score for two scanpaths. We compare S_1 to all the other scanpaths for the same image and then, take the average. Each scanpath for the image is compared to the rest of the related scanpaths in the same way. This yields 15 image-scanpath variation scores, whose mean corresponds to the gaze variation score of a single image. The higher this score is, the more variation exists in the gaze modality. We obtain a range of gaze variation scores for the whole set ($mean = 24.00$, $min = 11.22, max = 38.79$).

4.5 Correlation between Variations

In the previous subsections, we have quantified the variation in speech onsets, starting points, descriptions and gaze per image, see Appendix F for the images with the minimum and maximum scores across our variables of interest. We now turn to the correlation between the variation types. Since the initial common point is the image itself, we hypothesize that image features contribute to varying levels of variation in different modalities. We run Spearman's correlation between each type of variation.[8] When interpreting the magnitudes of the correlation coefficients, we use the terminology suggested by Prion and Haerling (2014). See Figure 3 for all correlation results.

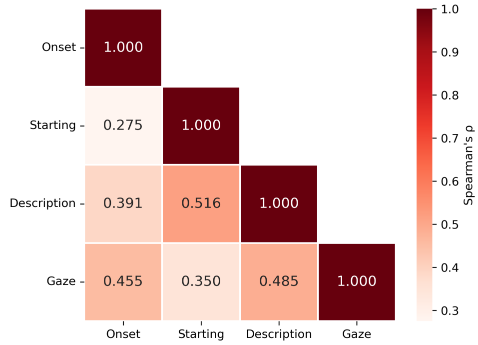


Figure 3: Spearman's correlation coefficients between the mean onsets per image (Onset), the variation in starting points (Starting), BLEU-2-based variation in full descriptions (Description), and the variation in gaze (Gaze) in the full dataset. Since higher BLEU scores mean less variation unlike the trends in the other measures, we utilize $1 - BLEU$ for better interpretability. All of the correlations are significant, $p < .001$.

We find a significant negative correlation, approaching moderate effect, between BLEU-2-based linguistic variation and the mean onset of an image (Spearman's $\rho = -0.391, p < .001$, see Appendix G for the regression line). This means that speakers start describing images that yield more similar descriptions earlier.[9] In addition, as starting points vary, image descriptions become less similar (moderate, Spearman's $\rho = -0.516, p < .001$), indicating that initial deviations continue until the end of language production.

We find that the variation in gaze significantly correlates with speech onsets (moderate, Spearman's $\rho = 0.455, p < .001$); the variation in starting points (weak, Spearman's $\rho = 0.350, p < .001$); and the variation in full descriptions (moderate, Spearman's $\rho = -0.485, p < .001$). These outcomes indicate that high variation in gaze tends to co-occur with longer onsets, high variation in starting points, and less similarity in descriptions.[10]

[7]This metric is similar to Self-BLEU (Zhu et al., 2018), which was proposed to calculate the diversity of the sentences generated by a model. In Self-BLEU, each generated sentence is compared to the rest of the generated sentences within a document, and an average of the whole set is computed to indicate how varied a model's generations are.

[8]We conduct Spearman's rank correlation analysis to uncover monotonic relations in the data. This type of correlation does not assume a particular distribution of the data (non-parametric, as opposed to Pearson's normality assumption). Since some of the signals we have investigated are non-normally distributed (e.g., speech onsets), and the dataset is relatively small, we opted for Spearman.

[9]Unlike this correlation, we find that speech onsets are not correlated with how many words or nouns are uttered.

[10]Investigating the correlation between these types of variation and the number of objects in an image is not straightfor-

The correlations reveal a connection between the variation in visual and linguistic modalities. We hypothesize that the underlying reasons for such variation partly reside in the features of an image, echoing the claims by Jas and Parikh (2015) and Berger et al. (2023). In this sense, similar images are expected to elicit similar amounts of variation. Hence, the results motivate our research into whether image features as encoded by pretrained models can capture the variation in gaze and language.

5 Similarity-based Prediction

In light of the correlation findings in Section 4, we expect image features to be predictive of the variation in visuo-linguistic signals to some extent. We explore if the similarity scores between image features encoded by pretrained models would be meaningful when capturing variation in human signals. In particular, we hypothesize that the signals that are more internal to the pretrained models' training objectives would be captured better. For instance, CLIP was trained with respect to an image-to-text alignment objective (Radford et al., 2021); hence, it would be reasonable to expect that signals that are more inherent to the visual and language data could be encoded better compared to speech onsets, which are never seen by the model.

Approach We employ an approach that was proposed as an alternative to training regression models and representational similarity analysis, for predicting fMRI signals given linguistic input (Anderson et al., 2016). Using the similarities between model-encoded stimuli (embeddings of concepts) and the corresponding fMRI responses, the authors predict the fMRI signals for novel stimuli for which embeddings exist. This approach has been utilized to assess the extent to which deep neural networks capture brain representations in language-only and visually grounded setups (Anderson et al., 2017; Bruera et al., 2023; Bruera and Poesio, 2023). We explain how we operationalize this extrapolation method for our purposes in Section 5.1. As this approach does not require training, it is suitable for shedding light on the predictive power of pretrained image representations, given the small size of the dataset we use. We determine the splits based on the images. Hence, to mitigate imbalance issues, we create 50 random split setups with ~90% training (277 images) and ~10% test sets (30 images), and report results on the average of these 50 setups. Across setups, the training sets have similar representative powers in terms of their CLIP vector similarities to the images in the corresponding test sets.

Visual encoders To encode the images, we exploit three visual encoders: CLIP, ViT, and a randomly initialized CLIP model (without training at all). We use the ViT-B/32 version of CLIP's visual encoder (Radford et al., 2021), and extract the final 512-dimensional output for each image. Since this encoder has been trained in coordination with CLIP's textual encoder (Radford et al., 2021), we expect it to capture not only vision-related features, but also properties that are aligned with language. In addition, we test the representations of a purely visual encoder trained on object recognition, ViT (Dosovitskiy et al., 2021). We extract the last hidden states from ViT, and use the vector corresponding to the `[CLS]` token as the image representation. Finally, we also experiment with a randomly-initialized version of CLIP (RNDCLIP), along the lines of what Berger et al. (2023) did to avoid the information learned during pretraining.

5.1 Predicting the Variation in Descriptions

From the training set, we retrieve k images that are closest to the target image—the image for which we predict a signal variation score—based on their representational similarities, echoing the k-nearest neighbors algorithm. The final score is the weighted average of the variation found in the neighboring images. The weights correspond to the similarity scores between the retrieved images and the target image.

As depicted in Table 1, we find significant, yet weak, positive correlations for almost half of the 50 split configurations both for CLIP and ViT, with no meaningful correlations for RNDCLIP. CLIP slightly outperforms ViT, suggesting that language alignment in the visual modality yields a potential benefit in estimating the variation in descriptions.

The loss corresponds to the average difference between the predicted and target scores across the dataset. The losses are similar across encoder types despite the differences in correlations. Since this method makes predictions based on the ground truth outputs of the retrieved set, it is likely that the predictions remain in a similar range.

ward, as current object detection algorithms annotate images exhaustively, yielding a high number for many images.

Model	Coefficient	Sig.	Loss
CLIP	**0.3380**	**27**	0.0738
ViT	0.3135	23	**0.0723**
RNDCLIP	0.0472	3	0.0744

Table 1: Predicting variation in descriptions with the similarity-based approach, $k = 277$. Averages over 50 random splits. 'Coefficient' and 'Sig.' correspond to Spearman's ρ correlation coefficient and how many runs out of 50 yield significant correlations with $p < 0.05$.

5.2 Predicting Onset

We perform the similarity-based prediction approach outlined in Section 5.1 to predict mean speech onsets per image. Since longer onsets can be associated with more cognitively demanding images, we are interested in the average onset elicited by each image. The results (see Table 2) indicate that, by using a larger sample of CLIP-encoded images, we can obtain predictions weakly correlating with the target onsets. The differences in the results when using different k values suggest that the choice of the retrieval set limits the boundaries of the predictions, even though the median image similarity score for $k = 1$ is 0.77 in the dataset.

Model	Coefficient	Sig.	Loss	Range
CLIP-277	**0.2981**	**18**	0.8216	3.37 - 3.50
CLIP-10	0.2500	10	**0.7989**	2.60 - 4.37
CLIP-5	0.2265	14	0.8149	2.26 - 4.81
CLIP-1	0.0640	4	1.0746	1.69 - 6.39
ViT	0.2428	17	0.8072	3.11 - 3.67
RNDCLIP	0.0350	3	0.8249	3.38 - 3.47

Table 2: Predicting mean speech onsets with the similarity-based approach. The numbers in the model names correspond to k when retrieving closest images from the training set. RNDCLIP and ViT with $k = 277$. 'Range' is the range of the predictions for the test set.

When we use 277 images encoded with ViT to obtain the image similarities, the correlation is weaker than the same setup with CLIP. When we encode the images with RNDCLIP, although the loss is quite similar to the other setups, there is no meaningful correlation. The predictions in general center around the mean onset, as they are based on the outputs from the retrieval set.

5.3 Predicting Starting Points

We utilize the similarity-based prediction algorithm to predict the first uttered nouns of the descriptions. Since this is a subtask of generating descriptions, we consider this an interesting use case. For each image, we represent the most common first noun as a one-hot vector (with the dimensions being 739, corresponding to the size of the first-noun vocabulary of the whole dataset). We report the accuracy of predicting the correct starting point.

Model	$k = 277$	$k = 10$
CLIP	13.00%	**31.73**%
ViT	**26.47**%	30.53%
RNDCLIP	11.27%	10.40%
Baseline - Random	4.00%	4.00%
Baseline - Most common	11.27%	11.27%

Table 3: Predicting starting points with the similarity-based approach and the baselines, percentage of correctly identified starting points for different k values.

As illustrated in Table 3, all setups attain scores that outperform the baseline where we predict random starting points (theoretically, for a uniform distribution of starting points, $1/739 = 0.14\%$). We also predict the most common starting point ('man'), which performs similarly to RNDCLIP. With pretrained encoders, it is better to utilize lower k to attain better accuracy, since very similar images likely contain similar objects that are mentioned earlier in the utterances. Both CLIP and ViT show similar performances when $k = 10$, hinting at the relation between their training objectives and starting points, which often correspond to the most salient entity in the image.

5.4 Predicting the Variation in Gaze

We apply the similarity-based approach to predict the variation in gaze. The results (Table 4) reveal that the gaze variation can be approximated to a moderate extent with CLIP. Using a smaller retrieval set is beneficial, suggesting a strong link between image properties and the variation in gaze. Since CLIP has a powerful visual encoder (Shen et al., 2022), it is reasonable that the similarities between image features encoded by CLIP seem more meaningful when predicting the variation in gaze.

The outcomes are in line with our hypothesis that signals that could be considered more internal to the models' training objectives would be captured better, whereas external signals can be captured weakly. For instance, speech onsets and surface form variation in descriptions can be deemed external to CLIP's space. Therefore, we claim that there could be room for incorporating such exter-

Model	Coefficient	Sig.	Loss	Range
CLIP-277	0.4035	30	4.0200	23.55 - 24.45
CLIP-10	0.4253	35	3.5774	17.05 - 29.63
CLIP-5	0.4435	33	**3.5707**	15.43 - 32.92
CLIP-1	**0.4687**	**39**	3.8889	11.22 - 38.79
ViT	0.3801	28	3.8847	22.62 - 25.67
RNDCLIP	0.0109	2	4.0571	23.76 - 24.26

Table 4: Predicting gaze variation using the similarity-based approach. Targets range between 11.22 and 38.79.

nal signals when training or fine-tuning pretrained multimodal models, and the models would benefit from such signals. It should be noted, though, since human processes are complex, there could be extraneous factors beyond image features that influence variation, which makes it difficult for models to capture these signals perfectly.

Min: 3.381 Max: 3.488

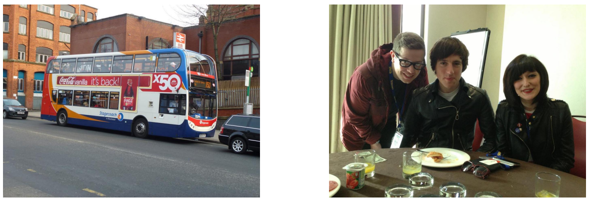
Figure 4: Images with the minimum and maximum predicted mean onsets. The image with the minimum was also predicted to elicit the lowest variation in gaze.

5.5 Examples

We illustrate the images with the minimum and maximum mean onsets as predicted by the similarity-based approach in Figure 4. Figure 5 depicts predicted variation in descriptions, and Figure 6 the predicted variation in gaze. We see a tendency to predict shorter speech onsets, more similar descriptions and gaze patterns in images containing a couple of people compared to scenes of streets with no visible or salient humans, a finding resonating with the conclusions drawn by Jas and Parikh (2015). This is potentially due to the salient and non-ambiguous nature of humans in images, as opposed to general street scenes with cars, buses and non-salient humans.

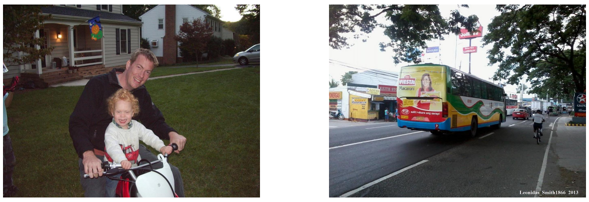
6 Conclusion

We quantified the variation in speech onsets, starting points, descriptions and gaze using a Dutch dataset of image descriptions with eye-tracking data. Our findings revealed the extent of variation in the process of describing images, and that variations in different signals correlate with each other. Furthermore, using a similarity-based prediction approach, we showed that image representations encoded by pretrained vision encoders capture variation in visuo-linguistic behavior to a weak-to-moderate extent. This pattern can be interpreted in light of models' pretraining objectives, as the predictions correlated more strongly for signals more internal to the objectives. Our study has implications for how human processes unfold as well as pretrained models' capabilities to represent such processes.



Min: 0.529 Max: 0.541

Figure 5: BLEU-2-based linguistic variation scores as predicted by the similarity-based approach.



Min: 23.666 Max: 24.308

Figure 6: Variation in gaze as predicted by the similarity-based approach.

Human and machine processing have differences, and we are motivated by the potential benefits of making the models increasingly knowledgeable about the multimodal landscape of human data. Although the impact of fine-tuning an already powerful pretrained model on a small-scale dataset with human signals could be modest, we hope that our work motivates the collection of more signals during crowdsourcing. For instance, it would be beneficial to take into account how long it took participants to complete a task given a certain stimulus, indicating the relative complexity and the uncertainty induced by the task as well as the stimulus. By inducing biases based on human signals, models can further take advantage of the information contained within such signals. Although it would be difficult to capture the full extent of the intricacies of human processing, this could help, for instance, a model interacting with human users to generate responses more aligned with human expectations.

Limitations

In this work, we use a dataset in Dutch; however, the crossmodal interaction between vision and language could show some variation based on the properties of the languages (i.e. word order and morphological constraints), leading to variation in visual attention and structural choices (Norcliffe and Konopka, 2015; Myachykov et al., 2011). Therefore, the findings might differ based on the languages of the datasets and the pretrained models. It would also be informative to explore other models and tasks, as well as explicit, discrete features that would contribute to the prediction of visuo-linguistic variation. Regarding the data, there could be possible noise in human signals and our preprocessing steps that affect the findings. Investigating the variation in gaze before/after speech onset with participant-specific analyses could also reveal interesting dynamics. As the dataset contains descriptions from 45 participants, with on average 15 participants describing each image, a different pool of participants (in particular, of a different size) may produce disparate results. A larger corpus may also allow for the training and fine-tuning of models. This is a line of work we have not explored in detail in this work, as a probing approach where we trained lightweight layers on top of image representations yielded even lower correlation coefficients and higher losses.

Ethics Statement

The data we use in this work had been collected following ethical guidelines (van Miltenburg et al., 2018b). Predicting or using eye-tracking of humans in the wild would require ethical approval, a line of research which we do not investigate, as our focus is on whether pretrained model representations can account for variations in human outputs. Since we use large pretrained models in frozen form, they may be perpetuating biases that are not desirable.

Acknowledgements

We would like to thank the members of the Dialogue Modelling Group for their valuable feedback regarding the project, and Andrea Bruera for his pointers to the work by Anderson et al. (2016). We also would like to thank the anonymous ARR reviewers for their thoughtful and useful feedback. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Özge Alacam, Eugen Ruppert, Ganeshan Malhotra, Chris Biemann, and Sina Zarrieß. 2022. Modeling referential gaze in task-oriented settings of varying referential complexity. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 197–210, Online only. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrew James Anderson, Benjamin D. Zinszer, and Rajeev D.S. Raizada. 2016. Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association* *for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Uri Berger, Lea Frermann, Gabriel Stanovsky, and Omri Abend. 2023. A large-scale multilingual study of visual constraints on linguistic selection of descriptions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2285–2299, Dubrovnik, Croatia. Association for Computational Linguistics.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language.

Kathryn Bock, David E. Irwin, and Douglas J. Davidson. 2004. Putting first things first. *The interface of language, vision, and action: Eye movements and the visual world*, pages 249–278.

Andrea Bruera and Massimo Poesio. 2023. Family lexicon: using language models to encode memories of personally familiar and famous people and places in the brain. *bioRxiv*.

Andrea Bruera, Yuan Tao, Andrew Anderson, Derya Çokal, Janosch Haber, and Massimo Poesio. 2023. Modeling brain representations of words' concreteness in context using gpt-2 and human ratings. *Cognitive Science*, 47(12):e13388.

Guy Thomas Buswell. 1935. *How people look at pictures: A study of the psychology and perception in art*. University of Chicago Press.

Monica S. Castelhano, Michael L. Mack, and John M. Henderson. 2009. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):6–6.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040.

Moreno I. Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, 36(7):1204–1223.

Moreno I. Coco and Frank Keller. 2015. Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive processing*, 16(2):131—150.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Xiao Ding, Bowen Chen, Li Du, Bing Qin, and Ting Liu. 2022. CogBERT: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3210–3225, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sibo Dong, Justin Goldstein, and Grace Hui Yang. 2022. Gazby: Gaze-based bert model to incorporate human attention in neural information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, page 182–192, New York, NY, USA. Association for Computing Machinery.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yulia Esaulova, Martina Penke, and Sarah Dolscheid. 2019. Describing events: Changes in eye movements and language production due to visual and conceptual properties of scenes. *Frontiers in Psychology*, 10.

Fernanda Ferreira and Gwendolyn Rehrig. 2019. Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, 34(9):1129–1139.

Albert Gatt, Emiel Krahmer, Kees van Deemter, and Roger P.G. van Gompel. 2017. Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, 41(S6):1457–1492.

Spandana Gella and Frank Keller. 2018. An evaluation of image-based verb prediction models against human eye-tracking data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 758–763, New Orleans, Louisiana. Association for Computational Linguistics.

Lila R. Gleitman, David January, Rebecca Nappa, and John C. Trueswell. 2007. On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4):544–569.

Zenzi M. Griffin. 2004. Why look? reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira, editor, *The interface of language, vision, and action: Eye movements and the visual world*, chapter 7, pages 213–247. Psychology Press, New York.

Zenzi M. Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11:274–9.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2023. What's in a name? a large-scale computational study on how competition between names affects naming variation. *Journal of Memory and Language*, 133:104459.

Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human attention in image captioning: Dataset and analysis. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8528–8537.

John M. Henderson. 2017. Gaze control as prediction. *Trends in Cognitive Sciences*, 21(1):15–23.

John M. Henderson and Fernanda Ferreira. 2013. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Taylor & Francis.

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021a. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021b. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Mainak Jas and Devi Parikh. 2015. Image specificity. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2727–2736.

Daniel Kahneman. 2012. *Thinking, Fast and Slow*. Penguin Books.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Willem J. M. Levelt. 1981. The speaker's linearization problem. *Philosophical Transactions of the Royal Society B*, 295:305–315.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Brian MacWhinney. 1977. Starting points. *Language*, 53(1):152–168.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Antje S. Meyer. 2004. The use of eye tracking in studies of sentence generation. In J. M. Henderson and F. Ferreira, editors, *The interface of language, vision, and action: Eye movements and the visual world*, chapter 6, pages 191–212. Psychology Press, New York.

Antje S. Meyer and Femke van der Meulen. 2000. Phonological priming effects on speech onset latencies and viewing times in object naming. *Psychonomic Bulletin & Review*, 7:314–319.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-Tracking*, 1st edition. Springer Publishing Company, Incorporated.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Andriy Myachykov, Dominic Thompson, Christoph Scheepers, and Simon Garrod. 2011. Visual attention and structural choice in sentence production across languages. *Language and Linguistics Compass*, 5(2):95–107.

Elisabeth Norcliffe and Agnieszka E. Konopka. 2015. *Vision and Language in Cross-Linguistic Research on Sentence Production*, pages 77–96. Springer India, New Delhi.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Aude Oliva. 2005. Gist of the scene. In *Neurobiology of attention*, pages 251–256. Elsevier.

Aude Oliva and Antonio Torralba. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Charlotte Pouw, Nora Hollenstein, and Lisa Beinborn. 2023. Cross-lingual transfer of cognitive processing complexity. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 655–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Susan K. Prion and Katie Anne Haerling. 2014. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10:535–536.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. 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*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yuqi Ren and Deyi Xiong. 2021. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3758–3769, Online. Association for Computational Linguistics.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, page 71–78, New York, NY, USA. Association for Computing Machinery.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.

Dan I. Slobin. 2003. Language and Thought Online: Cognitive Consequences of Linguistic Relativity. In *Language in Mind: Advances in the Study of Language and Thought*, pages 157–192. The MIT Press.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2647–2657.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 27–43, Online. Association for Computational Linguistics.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*.

Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm. 2018. SNAG: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–137, Melbourne, Australia. Association for Computational Linguistics.

Femke F. van der Meulen, Antje S. Meyer, and Willem J. M. Levelt. 2001. Eye movements during the production of nouns and pronouns. *Memory & Cognition*, 29:512–521.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018a. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. 2018b. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3658–3669. Association for Computational Linguistics.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

P.C. Wason and J.ST.B.T. Evans. 1974. Dual processes in reasoning? *Cognition*, 3(2):141–154.

Alfred L. Yarbus. 1967. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.

Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Studying relationships between human gaze, description, and computer vision. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746.

Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020. Human gaze assisted artificial intelligence: A review. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4951–4958. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

A Data Preprocessing

We use spaCy to extract the first noun of each description. The numbers of errors in terms of lemmatization and POS-tagging are as follows when using the small, medium, and large spaCy models for Dutch, respectively: 33, 32, 23 mistakes in the full descriptions, and 3, 2, 2 for the first nouns. As the utterances sometimes contain incomplete sentences and disfluencies, POS-tagging may not be reliable in such cases, especially in the later parts of the utterances. However, the large model was reliable both for full descriptions and the first nouns. Hence, we chose to use the data processed by the large model. The model was not able to tag any nouns in 7 descriptions; for those, we use the <unk> token as a placeholder starting point. We also skipped nouns such as 'photo' ('a photo of a car'), 'number' (as in 'a number of cats'), 'couple' (as in 'a couple of kids').

B Distribution of Speech Onsets

The histograms of the mean speech onsets and their standard deviations reveal non-normal distributions, as illustrated in Figure 7.

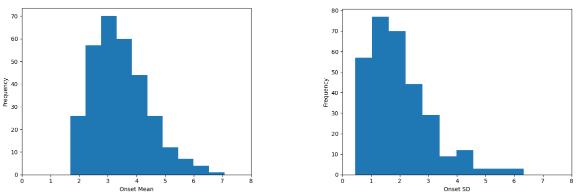


Figure 7: Distributions of onset means and SDs for the images in the whole dataset.

C Participant-Based Correlation Analysis

To have a better understanding of speaker-specific dynamics, in addition to calculating statistics per image, we also look into per-participant statistics.

Each participant describes around 100 images, each with a possibly different speech onset. We calculate the correlation between a participant's speech onsets and the BLEU-2-based linguistic variation score of the corresponding images. In 24 out of 45 participants, we find significant moderate negative correlations. All 45 participants have negative correlation coefficients, indicating that all participants tend to start describing an image earlier if that image elicits less linguistic variation across speakers. This suggests that although there can be speaker-specific and contextual factors, the features of an image can also have an overarching effect on the behavioral responses across speakers, and may allow for the prediction of such responses.

D BERTje-based Variation in Descriptions

We inspect linguistic variation by comparing the representations of the descriptions extracted using a Dutch BERT model (BERTje; de Vries et al., 2019). To calculate variation based on BERTje, we utilize the last hidden state corresponding to the [CLS] token for each description as the representation. Then, for each image, we calculate the pairwise cosine similarities between these representations. The average of these similarities is assigned as the variation found in the descriptions of an image. This method yields scores in the narrow range of $0.69 - 0.86$, which indicates semantically quite similar descriptions. Since most descriptions have semantics suitable for the corresponding image, the variation in the semantic space is not substantial. Between BERTje-based variation and speech onsets, we reveal a slight negative correlation (Spearman's $\rho = -0.212, p < 0.01$). The SD of speech onsets is even less correlated with BERTje-based variation (Spearman's $\rho = -0.151, p < 0.01$).

E Further Analyses on Linguistic Variation Metrics

We also combine BERTje- and BLEU-2-based variation scores by taking their mean. This metric yields correlations comparable to the ones achieved by the BLEU-2 version, with a moderate increase in the correlation to the starting point variation and mean onset, yet a decrease in the correlation to gaze variation. For the sake of simplicity, we opt for the BLEU-2 version.

We also compare the BLEU-2-based metric against human evaluations for a different dataset provided by Jas and Parikh (2015), which achieves a significant correlation (Spearman's $\rho = -0.40, p < .001$), albeit to a moderate extent. Jas and Parikh (2015) propose a metric that achieves a stronger correlation ($\rho = 0.72$). Note that the provided human annotations were obtained through 3 annotators evaluating sentence similarities without looking at the images (comparing only 2 sentences at a time). In our dataset, using our metric, we compare 1 description against 14. As a result, the procedure for human annotations may not be well-aligned with our method (i.e., our metric compares 1 sentence against 4 for their dataset, as each image has 5 descriptions).

F Example Images with Scores

We illustrate the images with the minimum and maximum scores per variable of interest as calculated with our metrics. Figure 9 depicts variation in the descriptions Figure 10 the variation in starting points, and Figure 11 the variation in gaze.

G Correlation between Human Signals of Variation

We illustrate the correlation between the mean onset and the BLEU-2 scores of full descriptions in Figure 8.

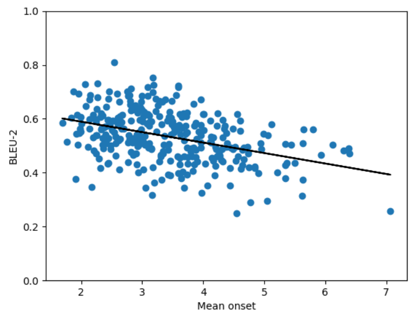


Figure 8: Correlation between mean onset and BLEU-2.

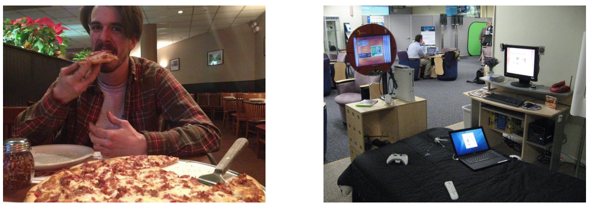
Min: 0.248

Max: 0.811

Figure 9: BLEU-2-based linguistic variation scores.

Min: 1 Max: 13

Figure 10: Variation in the number of unique starting points. For the image with the minimum score, all the speakers start with *keuken*, meaning kitchen. The image with the maximum score has descriptions starting with a variety of words: bureau, fitness, huiskamer, springding, atletiek, balk, hoek, tafel, plek, turnattribuut, restaurant, bank, turnobject.



Min: 11.22 Max: 38.79

Figure 11: Variation in gaze. The image with the minimum score elicited more similar scanpaths across speakers than the one with the maximum score.